

版权注意事项：

- 1、书籍版权归作者和出版社所有
- 2、本PDF仅限用于个人获取知识，进行私底下的知识交流
- 3、PDF获得者不得在互联网上以任何目的进行传播
- 4、如觉得书籍内容很赞，请购买正版实体书，支持作者
- 5、请于下载PDF后24小时内删除本PDF。

大数据思维与 应用攻略

王崇骏

编著

现象及感性思辨——溯源大数据
技术及选型思路——剖析大数据
实施及理性思考——实践大数据
机遇及应用思索——拥抱大数据



机械工业出版社
China Machine Press

内容简介

“大数据”是近年来最为热门的技术名词和概念名词之一，从其诞生之日起，就引起了“政产学研商用”各界的普遍热议，也获得了哲学家、科学家、技术研究者、工程研发人员等的普遍关注。

本书的整体行文是基于“说些历史、讲些故事、聊些技术、谈些思考”这样的思路展开的。本书共有13章，逻辑上分为四篇：

第一篇“现象及感性思辨”尝试对“数觉→数→数据→大数据”的历史脉络进行梳理并陈述社会各界迎接和拥抱“大数据”的若干事实。

第二篇“技术及选型思路”尝试从技术实现和部署实施的角度厘清大数据技术流程，并从多个视角阐述各环节面临的挑战和响应策略。

第三篇“实施及理性思考”尝试从管理策略、价值实现及思维方式三个角度厘清大数据落地应用涉及的技术和非技术问题。

第四篇“机遇及应用思索”尝试在对“互联网+”技术发展脉络及国际经济形势进行梳理的基础上，研判大数据潜在的发展机遇和应用场景。

本书的初始行文动机是“归纳所见所闻、总结项目经历、独立自主思考”，希望从独立、客观的第三方视角介绍和分析“大数据”及“大数据”相关的技术观点、执行思路和关键问题。期望本书能够为大数据相关工作者、普通院校大数据相关专业的研究生和本科生以及对大数据有兴趣的读者等提供一些借鉴。

大数据思维与

应用攻略

王崇骏 编著



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

大数据思维与应用攻略 / 王崇骏编著. —北京: 机械工业出版社, 2016.7

ISBN 978-7-111-54261-2

I. 大… II. 王… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字 (2016) 第 157469 号

大数据思维与应用攻略

出版发行: 机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码: 100037)

责任编辑: 余 洁

责任校对: 殷 虹

印 刷: 北京市荣盛彩色印刷有限公司

版 次: 2016 年 7 月第 1 版第 1 次印刷

开 本: 186mm×240mm 1/16

印 张: 25.5

书 号: ISBN 978-7-111-54261-2

定 价: 59.00 元

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

客服热线: (010) 88379426 88361066

投稿热线: (010) 88379604

购书热线: (010) 68326294 88379649 68995259

读者信箱: hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

Preface 序

或许还从来没有一个技术名词像“大数据”一样，从其诞生之日起，就得到了“政产学研商用”的一致认同，并引起了哲学家、科学家、技术研究者、工程研发人员的普遍关注。大数据能够受到如此重视的原因可能是：不同的利益主体及不同的工作群体都不约而同地意识到了由“大数据”引发和驱动的思维变革、问题凝练、技术挑战和价值发现。

作为一名大学全职教师，在教学和科研的过程中，我有幸接触、参与和主持了“大数据”相关的课程讲授、项目研发和咨询服务。在教学相长及每一个项目场景的迭代研发过程中，我对“大数据”及数据驱动的项目研发颇有感触。当我尝试把这种感触和心得与学生、同事、学界小同行及项目甲方（有的是软件企业，有的是直接应用方，比如政府、企事业单位等）进行交流的时候，我发现针对不同特点的受众，必须有意地调整相关说辞，否则很难有理论技术及应用情怀方面的共鸣和共振。我想这其中的原因最有可能在于：

1) 虽然大家都认为“大数据”是有价值的，但因为价值是极具主观色彩的，所以针对同一个应用场景，不同角色的利益主体出于自身的价值观和心理习惯会持有不同的价值期望。这就意味着，不在同一个价值观体系里的所有讨论，其成效大多会大打折扣。

2) 拥抱“大数据”是一种涉及多个专业领域和工种的集体协作行为，具有不同知识背景和方法论的不同个体如何有效地进行协作本身就是一个难题，需要在多边理解与互补的基础上互融互通、相互成就。

出于对“大数据”的敬畏，本书尝试将笔者面向不同群体陈述和演讲的观点加以梳理和规整，希望从不同的视角和层面分析和研判“大数据”这件事。但是“大数据”涉及的内容太多，笔者无法也从来没有想过，仅仅通过一本书就能详尽介绍其方方面面。本书尝试从相对宏观的角度对“大数据”进行介绍，希望通过这种介绍，能够让大家对“大数据”形成初步的印象，而具体的细节则需要通过其他途径深入了解。

本书的整体行文是基于“说些历史、讲些故事、聊些技术、谈些思考”这样的思路展开的。本书共有13章，逻辑上分为四篇，分别是：

第一篇 现象及感性思辨：本篇尝试对“数觉→数→数据→大数据”的历史脉络进行梳理，并通过社会各界迎接和拥抱“大数据”的若干事实，厘清几个基本问题：“大数据”是什么？为什么会有“大数据”？“大数据”得到各界关注的原因是什么？社会各界或各个利益主体是如何拥抱“大数据”的？本篇包括3个主要章节，分别是：

□ 第1章 大数据溯源

□ 第2章 大数据现象

□ 第3章 大数据产业

第二篇 技术及选型思路：本篇尝试从技术实现和部署、实施的角度厘清大数据技术流程，并从多个视角和层面阐述各个环节所面临的挑战和机遇，重点叙述不同知识背景的研究群体针对大数据的态度、行动和思维方式。本篇将通过对“数据采集→数据存取→数据建模→知识发现”技术流程的梳理，以及各个环节技术选型的应用提示，厘清几个基本的技术问题：从何处及如何获得数据？将数据存储在哪里及如何管理？如何分析数据以探明数据背后的知识和洞见？本篇包括6个主要章节，分别是：

□ 第4章 大数据支撑技术

□ 第5章 数据采集与整合

□ 第6章 数据存储与管理

□ 第7章 数据表示与理解

□ 第8章 数据理解与建模

□ 第9章 知识发现与应用

第三篇 实施及理性思考：本篇尝试从管理策略、价值实现及思维方式三个角度厘清大数据落地应用所涉及的技术和非技术问题，并从多个视角和层面梳理各个环节的要点和细则。此外，本篇将围绕“大数据实施及过程管理→大数据价值及价值评估→大数据思维及价值实现”这一主线，给出各个环节的应用提示，并厘清几个基本问题：“大数据”的价值在哪里以及如何实现？如何部署、实施一个大数据项目？应该以怎样的思维方式和执行策略应对“大数据”的挑战？作为一个数据分析师应该具有哪些情怀？本篇包括3个主要章节，分别是：

□ 第10章 大数据实施

□ 第11章 大数据价值

□ 第12章 大数据思维

第四篇 机遇及应用思索：本篇在对互联网的技术发展脉络及国际经济形势进行梳理的基础上，分析了在“互联网+”概念被热炒及全民总动员的当代，“大数据”的潜在发展机遇和应用场景，并通过对电子商务、工业4.0、互联网金融这三条主线的扼要描述和分析，厘清几个基本问题：“互联网+”的本质是什么？究竟是“互联网+X”还是“X+互联网”？“互联网+商务”“互联网+工业”“互联网+金融”的本质、门类及潜在机遇有哪些？作为一个

数据分析师，在“互联网+”的环境下应该具有哪些情怀？本篇包括1个主要章节：

□ 第13章 大数据机遇

本书的每一章都围绕某个专题展开叙述，独立成文，读者可以根据自己的兴趣和时间选择性阅读。本书的行文主要有两种字体：以宋体行文的部分主要描述相关理论、技术，以及必要的分析和说明；以楷体行文的部分大多是基于上下文罗列的一些佐证、案例和思考；此外，本书还分别以显式和隐式的方式给出了笔者对于一些技术选型、场景分析及感悟心得的“应用提示”。

本书的初始行文动机是“归纳所见所闻、总结项目经历、独立自主思考”，希望从客观、独立的第三方视角介绍和分析“大数据”及与“大数据”相关的技术观点、执行思路和关键问题。在不影响阅读的情况下，本书对所涉及的公司产品的介绍简明扼要，更多详细信息可参见给出的推荐链接或参考读物。希望通过这样的章节安排及内容梳理，本书能够给“大数据”相关工作者一些参考，比如有意建设大数据项目的创业型公司创始人、企业主或政府部门的主管及信息主管（作为技术丛书和应用手册），有意向数据型公司转型的传统软件企业技术人员，包括市场人员、研发人员和主管（作为应用指南及技术白皮书），处于战略转型期的传统企业主或信息主管（作为技术丛书），大数据项目研发工程技术人员（作为技术丛书及应用指南），普通院校大数据相关专业的研究生和本科生（作为教辅材料），或对大数据有兴趣的读者（作为科普读物）等。

本书的编写及出版得益于诸多前辈、同仁和南京大学计算机科学与技术系及南京大学软件新技术国家重点实验室的各位领导、同事的提携、关心和鼓励。感谢各类项目的资助方以及我所在研究团队“南京大学智能信息处理研究组”的小伙伴，多年来共同努力和协作完成的一个个项目为本书的撰写提供了大量素材，同时也让我有更多的可能性去思考所有这些成功（当然也有失败的）案例背后的大数据逻辑。还要特别感谢“南京大学智能信息处理研究组”的吴骏博士、张雷博士及彭岳、徐鸣、陆恒杨、王楠、李明、王陆霞、夏丽、谭龙海、陈鹏飞、冯艺琳、唐驰、谢璐遥、李永春等同学，在对本书进行通本润色的过程中，他们给予了极大的帮助。本书行文伊始，我就将本书的编写计划、组织结构与南京大学的谢俊元教授、陈家骏教授、郑滔教授进行交流和沟通，三位教授均给予了很多务实的建议，在本书的整个编写过程中，几位教授也在不同的场合、时机关注本书的编写进度。这些对于笔者而言都是莫大的支持和鼓舞，在此一并感谢。本书的编写和出版还离不开机械工业出版社华章公司姚蕾编辑的建议和鼓舞，尤其是行文过程中几乎不间断的支持和鼓励，才使本书得以顺利完稿，再次表示感谢。

由于水平有限及知识面、价值观的狭隘，书中有疏漏和不足之处在所难免，敬请各位专家和读者批评指正。

目 录 Contents

序

第一篇 现象及感性思辨

第1章 大数据溯源 3

- 1.1 引言 3
- 1.2 数觉及数的起源 7
- 1.3 模拟与数字计算 10
- 1.4 从数据到大数据 15
- 1.5 大数据时代 19
- 1.6 本章小结 23
- 本章参考文献 23

第2章 大数据现象 25

- 2.1 引言 25
- 2.2 政界大数据 28
- 2.3 业界大数据 33
- 2.4 学界大数据 39
- 2.5 本章小结 44
- 本章参考文献 45

第3章 大数据产业 46

- 3.1 引言 46

3.2 大数据产业环境 49

- 3.2.1 政策环境 49
- 3.2.2 应用环境 51
- 3.2.3 技术环境 52

3.3 大数据产业地图 53

- 3.3.1 大数据产业地图由来 53
- 3.3.2 大数据产业地图明细 54
- 3.3.3 大数据产业地图意义 61

3.4 大数据应用提示 62

- 3.4.1 大数据中文解析及提示 62
- 3.4.2 大数据应用场景及策略 64
- 3.4.3 大数据陷阱及应用提示 65

3.5 本章小结 67

本章参考文献 68

第二篇 技术及选型思路

第4章 大数据支撑技术 71

- 4.1 引言 71
- 4.2 大数据流程 73
 - 4.2.1 显式挑战 74
 - 4.2.2 隐式困难 76
 - 4.2.3 评估思路 78

4.3 基础支撑技术	78
4.3.1 数据采集	79
4.3.2 数据存储	81
4.3.3 数据建模	82
4.3.4 计算架构	85
4.4 高级支撑技术	90
4.4.1 云计算背景	90
4.4.2 云计算定义	91
4.4.3 云计算本质	93
4.4.4 应用提示	96
4.5 本章小结	97
本章参考文献	98

第5章 数据采集与整合

5.1 引言	99
5.2 大数据的数据源	101
5.2.1 数据分布	101
5.2.2 内部数据	103
5.2.3 互联网数据	105
5.2.4 应用提示	105
5.3 内部数据及内部数据 采集	106
5.3.1 目标任务	106
5.3.2 关键技术	107
5.3.3 ETL 工具	110
5.3.4 应用提示	111
5.4 互联网数据及互联网 数据采集	113
5.4.1 目标任务	113
5.4.2 关键技术	114
5.4.3 开源网络爬虫	118
5.4.4 应用提示	120
5.5 本章小结	121
本章参考文献	123

第6章 数据存储与管理

6.1 引言	124
6.2 数据组织	127
6.2.1 集中与分布	128
6.2.2 SQL 与 NoSQL	130
6.3 数据存储	138
6.4 云存储	141
6.5 本章小结	144
本章参考文献	145

第7章 数据表示与理解

7.1 引言	146
7.2 度量方法	149
7.2.1 相似系数函数	150
7.2.2 距离函数	152
7.3 数据规范	154
7.4 特征工程	155
7.4.1 特征表示	156
7.4.2 特征提取	156
7.4.3 特征选择	175
7.5 应用提示	178
7.6 本章小结	181
本章参考文献	181

第8章 数据理解与建模

8.1 引言	183
8.2 机器学习	185
8.3 非监督学习	187
8.3.1 K-Means	188
8.3.2 EM	189
8.4 监督学习	192
8.4.1 回归	192
8.4.2 分类	196
8.5 本章小结	226

本章参考文献	227	10.3.1 生产流程管理	274
第9章 知识发现与应用	229	10.3.2 技术流程管理	277
9.1 引言	229	10.3.3 知识流程管理	279
9.2 从机器学习到数据挖掘	233	10.4 商务管理	282
9.2.1 统计与统计学	234	10.4.1 商业模式价值逻辑	282
9.2.2 智能与人工智能	235	10.4.2 大数据与商业模式	283
9.2.3 人工智能与机器学习	237	10.4.3 典型商业模式示例	287
9.2.4 数据挖掘及技术路径	239	10.5 本章小结	290
9.2.5 应用提示	245	本章参考文献	291
9.3 从数据挖掘到数据科学	246	第11章 大数据价值	292
9.3.1 从“惊奇”引发的科学 之母	246	11.1 引言	292
9.3.2 从“科学”引发的研究 范式	249	11.2 从数据到价值	294
9.3.3 从“数据”引发的数据 科学	251	11.2.1 数据的价值	295
9.4 从算法到大数据方法论	252	11.2.2 信息的价值	297
9.4.1 演绎与归纳	252	11.2.3 知识的价值	299
9.4.2 因果与相关	255	11.2.4 应用提示	300
9.4.3 定律与模型	257	11.3 从闭环到开环	302
9.5 本章小结	260	11.3.1 垂直应用价值	302
本章参考文献	260	11.3.2 平台集成价值	303
		11.3.3 生态协同价值	305
		11.3.4 应用提示	305
		11.4 大数据评估	306
		11.4.1 数据价值评估	306
		11.4.2 数据质量评估	310
		11.4.3 平台价值评估	312
		11.4.4 应用提示	315
		11.5 本章小结	321
		本章参考文献	322
第三篇 实施及理性思考		第12章 大数据思维	323
第10章 大数据实施	265	12.1 引言	323
10.1 引言	265	12.2 数据层	325
10.2 工程管理	267	12.2.1 数据全采样	325
10.2.1 思维层的应用模式梳理	267		
10.2.2 开发层的工程实施路径	270		
10.2.3 运维层的平台应用 保障	273		
10.3 技术管理	274		

12.2.2	数据交叉复用	327
12.2.3	数据云化存储	328
12.3	分析层	330
12.3.1	相关重于因果	330
12.3.2	效率重于精度	332
12.3.3	离线分析+实时运行 ...	334
12.4	应用层	336
12.4.1	数据质量溯源	336
12.4.2	服务和应用	340
12.4.3	开放和合作	342
12.5	本章小结	345
	本章参考文献	347

第四篇 机遇及应用思索

第13章	大数据机遇	351
13.1	引言	351
13.2	互联网+	356
13.3	电子商务	359
13.3.1	电子商务概述	359

13.3.2	移动电子商务	362
13.3.3	跨境电子商务	363
13.3.4	应用提示	365
13.4	工业互联网	368
13.4.1	基本概念	368
13.4.2	笑脸曲线	368
13.4.3	工业4.0	371
13.4.4	应用提示	376
13.5	互联网金融	380
13.5.1	基本概念	380
13.5.2	面向投融资的互联网 金融	381
13.5.3	面向支付的互联网 金融	384
13.5.4	其他类型的互联网 金融	387
13.5.5	应用提示	390
13.6	本章小结	392
	本章参考文献	394

跋	395
---------	-----

第一篇 Part 1

现象及感性思辨

1.1 引言

- 第1章 大数据溯源
- 第2章 大数据现象
- 第3章 大数据产业

科学技术的不断进步和人类需求的持续膨胀，就如两个互相咬合和彼此驱动的齿轮伴随着人类文明进程不断地发展：需求的膨胀不断刺激科学研究的持续进行，而科技的进步驱动着需求的进一步膨胀。目前社会各界普遍热议和关注“大数据”，恰是由于人们对数据价值的期望（需求）与目前对数据处理的研究和技术水平不匹配，并因为各界的普遍价值认同，“大数据”的概念被炒作得近乎神话。本篇尝试对“数觉→数→数据→大数据”历史脉络进行梳理，并通过社会各界迎接和拥抱“大数据”的若干事实，厘清几个基本的问题：“大数据”是什么？为什么会有“大数据”？“大数据”得到各界关注的缘由是什么？社会各界或者各个利益主体是如何拥抱“大数据”的？

本篇包括3章内容，分别是：

第1章 大数据溯源 尝试梳理在人类文明进程的不断发展过程中，因为人类需求的不断膨胀和科技的不断发展而引发的，或许是必然的“数觉→数据”及“数据→大数据”的历史脉络，并介绍了不同利益主体对“大数据”的理解和定义。

第2章 大数据现象 尝试梳理“大数据”这一概念得到“热炒”的当下，因为不同的利益驱动使然而引发“政产学研商用”各界迎接和拥抱大数据的各类行动举措和思维态度，并分析了大数据这一概念得到多边热议的动机和缘由。

第3章 大数据产业 尝试梳理“大数据”这一概念持续得到多边认同的当下，因为不同的价值期望使然而引发不同利益主体迎接和响应大数据带来的机遇和挑战的各类决策和行动，并给出大数据潜在应用场景的理性评判依据。

【关键字】 大数据，历史脉络，多边热议，产业环境

大数据溯源

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的杨骏元、汤北亮、王陆霞、李明、唐驰、王姗姗等几位同学的协助，在此表示深深的谢意。

1.1 引言

140 亿年前，宇宙诞生……

46 亿年前，地球诞生……

38 亿年前，简单生命体出现……

1500 万年至 1000 万年前，腊玛古猿出现……

400 万年至 100 万年前，南方古猿出现……

200 万年至 150 万年前，能人出现……

200 万年至 20 万年前，直立人出现……

20 万年至 1 万年前，智人出现……

以上一组简单的数据勾勒出人类在整个历史长河中的进化史轮廓。与动物仅仅通过遗传进化不同，人类在进化过程中发展和演化出了一种非遗传性的继承：通过独一无二且日益发达的文化媒介，将知识和传统留给后代。这种文化传统使得人类以很快的速度和加速度进化并最终成为这个地球的统治者，而遗传进化退居于次要的位置。

这里所说的知识，指的是人类在改造世界的实践中所获得的认识和经验的总结归纳，可以指导解决实践问题的观点、经验、程序等信息。

或许正因为人类有了可以把知识和传统传递给后代的文化媒介，所以通过本身遗传系统所传递的信息也就愈来愈少。动物有许多生存的本领是通过遗传系统直接传递给后代的，而人除了吃喝哭喊之外，绝大部分生存本领只有靠后天学习。因此，发现知识、传递知识和学习知识是人类文明进程中亘古不变的主题。

野中郁次郎和竹内广孝在1995年提出的SECI (Socialization, Externalization, Combination, Internalization) 模型专门阐述了知识构建和管理的完整过程:从噪声中分拣出数据,转化为信息,升级为知识,升华为智慧,让信息从庞大无序到分类有序。

数据是指描述事物的符号记录,是构成信息和知识的原始材料,如图形、声音、文字、数、字符和符号等;信息一般指数据所包含的意义,可以使数据所描述事件的不确定性减少。数据、信息和知识的关系可以描述为:数据是信息的载体,信息是知识的载体。知识可以从数据中发掘出来,即知识发现。

知识发现是从数据(库)中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的过程,是将低层数据转换为高层知识的过程。其中的一个重要步骤是数据挖掘,也有很多场合将知识发现与数据挖掘有意无意地混淆。数据挖掘一般是指从大量的数据中通过算法发现隐藏于其中的信息的过程。关于数据挖掘与知识发现的详细介绍,参见9.2.4节,此处不赘述。

数据挖掘通常与计算机科学有关,通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标,而机器学习是其中最为重要的基础,其基本的分析方法包括:聚类、分类、关联规则发现、预测等。机器学习是一门多领域交叉学科,专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。关于机器学习的详细介绍参见9.2.3节。

如上所述,作为信息的载体,数据是指描述事物的符号记录,鉴于其承担的功能,数据具有很多天生的特征(数据层特征),比如异构、分布、多态、多模式等。

1) 异构性。数据的异构性主要表现在系统异构、模式异构、来源异构三个方面:

①数据源所依赖的业务应用系统、数据库管理系统乃至操作系统之间的不同构成了系统异构性。

②模式异构指的是数据源在存储模式上的不同。存储模式主要包括关系模式、对象模式、对象关系模式和文档嵌套模式等,其中关系模式(关系数据库)为主流存储模式。同时,即便是同一类存储模式,它们的模式结构可能也存在着差异,不同的关系数据管理系统的数据类型等方面并不是完全一致的,如DB2、Oracle、Sybase、Informix、SQL Server、Foxpro等。

③来源异构,即内部数据源和外部数据源之间的异构。

2) 分布性。数据的分布性指的是数据的分布特征。一组数据的分布特征可以从以下三个方面进行测度:

①集中趋势测度,如众数、中位数、分位数、均值、几何平均数、截尾均值等。

②离散程度测度,如极差、内距、方差和标准差、离散系数等。

③偏态与峰度测度,如偏态及其测度、峰度及其测度等。

在某种意义上而言,正是由于数据具有分布性,我们才可以挖掘出数据内部的模式,并对数据进行分析以得到某种结果、做出某种策略。

3) 多态性。数据的多态性即数据的多样性。一般而言,数据可分为两类:结构化与非结构化数据。在信息社会,信息可以分为两大类:一类信息能够用数据或统一的结构加以表示,我们称之为结构化数据,如数字、符号;而另一类信息无法用数字或统一的结构表示,如文本、图像、声音、网页等,我们称之为非结构化数据。数据的多样性在现今世界表现得更加明显,由于数据采集与存储技术的不断进步,生活中所有的数据都可以保存下来,来源的广泛与用途的多样使得数据变得丰富而有价值。

4) 多模式性。数据模式指对某一类数据的结构、属性、联系和约束的描述。数据模式是基于选定的数据模型对数据进行“型”方面的刻画,而相应的“实例”则是对数据“值”方面的描述。先有数据模型,才能据其讨论相应数据模式,有了数据模式,就能依据该模式得到相应的实例。正是数据的多模式性,才能够通过构造复杂的数据结构来建立数据之间的内在联系与复杂关系,从而构成数据的全局结构模式。

除了上述数据自身的特点,针对具体的应用场景,还需要考虑其他一些来自数据层的特征,比如数据的质量、数据的活度、数据的厚度、数据的规模等。

1) 数据的质量:正所谓“失之毫厘,谬以千里”,数据是否具备可靠性和有效性,直接影响到数据分析过程以及是否能得出正确的结论并做出正确的决策。而在大数据时代,高质量的数据是大数据发挥效能的前提和基础,强大、先进的数据分析技术是大数据发挥效能的重要手段。对大数据进行有效分析的前提是必须要保证数据的质量,专业的数据分析工具只有在高质量的大数据环境中才能提取出隐含的、准确的、有用的信息,企业基于这些高质量分析结果所做出的各项决策才不至于偏离正常轨道;否则,即使数据分析工具再先进,在充满“垃圾”的大数据环境中也只能提取出毫无意义的“垃圾”信息。因此,数据质量在大数据环境下显得尤其重要。

2) 数据的活度:数据的活度即数据持续更新的频度,也称为数据的新鲜度。以银行借贷业务和手机通话记录为例,一般来说,我们的工资每月发放一次,而手机通话记录则每天每小时都会产生,详细地描述了一个人与其他人交往的情况,加上个人在互联网中娱乐、购物等记录,形成了一个生活轨迹的画像。一般数据活度越大,其可参考的价值可能就越高。比如电商平台在夏季向用户推荐商品时,使用用户最新(春夏)的购买数据所做的分析,显然比使用往年(春夏)的购买数据所做的分析更为准确和有意义。数据的这种特性依赖于人类的活动,如果用户发生某类行为的次数越频繁,则该类数据的活度也就越高,反之亦然。根据数据的活度进行分析,可以得出很多有意义的结论,比如根据数据活度将用户分类等。

3) 数据的厚度:数据的厚度指的是数据的维度大小以及每个维度的语义丰富程度。一般来讲,一条数据的厚度是不定的,也就是说不同领域中的数据厚度是不相同的,可能很大,也可能很小。如在统计领域中,一条具有完整意义的数据一般有时间、地域和指标三个维度。维度一般是越多越好(往往厚度越大),以一个用户为例,如果你仅仅知道他的姓名、住址、电话等信息,你对他的了解就很有限;如果你知道他的体育爱好(比如打篮球、打乒乓球等)、文化爱好(比如喜欢读史学作品等),你可以对他进行更加有针对性的营销(语义信息

更多),比如推荐 NBA 球星的球鞋、推荐《史记》和《全球通史》等书籍,这样成功的概率就大。当然在分析数据时,一味地追求高维度分析,也会使得分析结果变差,这是由于维度可能是噪声,或者维度之间存在相关性,或者维度与具体问题不相关。也就是说,数据的厚度不是单纯的维度问题,还包含与其蕴含的语义丰富度。

4) 数据的规模:数据一般分为结构化和非结构化数据,一般的数据库属于关系型数据库,以存储结构化数据为主,级别在 TB 级。随着数据,比如通话详单的积累,数据会越来越大,甚至达到 PB、EB、ZB 级。大数据区别于传统数据的一个显著特性是数据的规模,因此产生了如何进行存储、并行计算、挖掘数据内部模式等话题,进而形成了大数据概念,催生了大数据时代。

阿基米德说过:“给我一个支点,我能撬起地球。”仿照类似的语调,微软的史密斯这样说:“给我提供一些数据,我就能做一些改变。如果给我提供所有数据,我就能拯救世界。”现在看来,阿基米德的话语虽然很朴素,但是在那个年代能够提出,还是很振奋人心的。虽然我们没有生活在那个年代,不过后面的这句话,我们都是当事人,因为我们现在正生活在这样的大数据时代。

一般意义上,大数据是大到无法通过人工在合理时间内截取、管理、处理并整理成为人类所能解读的信息。此定义或许还不足以完全描述大数据,因此附加了一个普遍认可的 4V 特征:

1) 数据来源多、体量大,描述为“Volume”。这个特征隐含的应用提示是:在具体的部署实施过程中,数据存取必须支撑海量的数据并发访问,并且数据处理的性能必须支持海量吞吐率。

2) 数据来源广、类型多,描述为“Variety”。这个特征隐含的应用提示是:在具体的部署实施过程中,数据存取必须支持多格式、多模式的高效管理,以及数据分析手段必须支持多格式、多模式的有效分析等。

3) 数据来源速度快,描述为“Velocity”。这个特征隐含的应用提示是:在具体的部署实施过程中,数据采集速度要快、数据存取速度要快、数据分析速度要快,而所有的这些要求都对相关的技术选型、策略定位有很大的影响。

4) 数据有用、有价值,描述为“Value”。这个特征隐含的潜台词是:价值密度稀疏。这是因为数据的有用和有价值都是有目标指向的,确切地说,应该是数据对于某个具体的应用而言是有价值的。由于数据量巨大,因此对于某个具体的应用而言,海量的数据中可能只有一部分是有价值的。

显然上述的定义及 4V 特征的描述只是一个普适的通用描述,在具体的落地实施过程中,不同的公司及团体也会根据不同的价值观和方法论提出其他具体特征。

IBM 提出的真实性(“Veracity”)特征是从大数据部署实施过程中数据质量的维度考虑的,IBM 认为:真实性是当前企业亟待考虑的重要维度,将促使他们利用数据融合和先进的

数学方法进一步提升数据的质量,从而创造更高的价值。

引发人们对大数据关注的原因有许多,其中之一必定是数据有价值。如果说知识及其获取是一件有价值的事情的话,那么如何实现“数据→信息→知识”就是一个重要的环节。与这个技术流耦合的关键问题还有:数据在哪里,如何获得数据,知识如何体现在具体的应用中以获取其价值等。这些问题比较复杂,因为它们与具体的应用场景、应用模式和商业模式有直接的关系。

本章尝试解释一些更简单的问题,比如:人生来就有数据这个概念吗?数据是如何产生的?如何使用这些数据?为什么数据会变“大”,以及大数据得到如此关注的缘由在哪里?为了有效回答上述问题,本章尝试梳理在人类文明进程的不断发过程中,因为人类需求的不断膨胀和科技的不断发展而引发的或许是必然的“数觉→数据”及“数据→大数据”的历史脉络,并介绍不同利益主体对大数据的理解和定义,本章下面的结构安排如下:1.2节介绍人类还处于蒙昧阶段就已经具有的“数觉”以及作为地球主宰的生物种类如何从一开始仅有的“数觉”逐步产生出“数”的概念;1.3节介绍因为人类的聪明才智发明和创造的一些计算工具(包括模拟的和数字的),更具有变革意义的是,图灵机及全电子通用计算机的发明直接将人类推进到如今丰富多彩的信息化时代和本文关注的大数据时代;1.4节简单介绍数据的产生和催生大数据时代来临的若干技术和非技术缘由;1.5节简单叙述和回顾我们正生活的大数据时代。

1.2 数觉及数的起源

所谓数觉,指的是在一个小的集合里,增加或者减去一个元素的时候,尽管未曾直接知道增减,也能够辨认到其中有所变化。有研究表明若干种动物具有数觉,而人是否有数觉则是一个难以研究和回答的问题。

有个农场主计划打死一只在望楼里筑巢的乌鸦,试了多次,始终未成功。因为人一走近,乌鸦就离开了巢,飞开了,它栖在远远的树上守着,等到人离开了望楼,才肯飞回巢。这样的试验继续,两个人走进望楼,一个人留着,一个人出来走开了,但是乌鸦并不上当,它等着直到留在望楼里的人也走了出来才罢。试验一连做了几天:两个人,三个人,四个人,都没有成功,五个人的时候成功了。五个人首先都进了望楼,留一个在里面,其他四人走出来,离开了。这次乌鸦却数不清了,它不能辨别四与五,马上就飞回巢了。

实际上,我们所知道的最惊人的例子要算一种叫作“独居蜂”(solitary wasp)的昆虫。这种母蜂在每个巢里下一个卵,并且在巢里面预先储藏了一批活的尺蠖,作为幼虫孵化后的食料。使人吃惊的是,各类独居蜂在每个巢里所放的尺蠖数目都是一定的:有些放五条,有些放十二条,多的甚至于有放二十四条的。最特别的是一种叫作“螺赢”的蜂,这种蜂雄的比雌的小得多。母蜂能用神秘的方法辨别孵化出来的幼虫是雄的还是雌的,并且据此相应地分

配食品的数量；它并不改变捕获物的大小和种类，只是给雄卵存储五条尺蠖，给雌卵存储十条尺蠖。

类似上述的一些例子，在很多动物上都可以做一些有趣的实验得到验证，但是，“人是否有数觉”却是一个难以回答的问题，既无法证明，也无法用实验论证。这是因为人总有意无意地用其他技能来帮助他的直接数觉（这或许正是使人类成为地球主宰的一种智慧）。但是，也有心理学的实验表明：未经过专门技术培训的普通文明人，其直接视觉数觉很少能超过四。或许也能从一些文字痕迹中发现一些端倪，比如英文的 thrice 和拉丁文的 ter，都具有同样的双重意义：三倍和许多；再比如拉丁文的 tres（三）和 trans（超过）、法文的 tres（极，非常）和 trois（三）等。

其实，不管人类是否有数觉或者数觉的大小是多少，后续逐步发展起来的数的概念、计数、算术逐步把人类向新的文明阶段推进。

石子计数、结绳计数、刻痕计数（龟壳、木片、竹片）是最经典的三类计数方式。在拉丁文中，“计算”一词写作“Calculus”，本意即为计算用的石子。用天然石子计数，是人类早期常用的一种计算方法，是每个民族都经历过的历史阶段。石块、木块等物虽然能计数，可是不太“保险”，稍不留意，一脚碰着就乱了套。在文字发明前，结绳计数几乎是世界各地的人们都曾使用过的一种记事方法，《易九家言》就有“事大，大结其绳；事小，小结其绳，结之多少，随物众寡”的说法，而《易·系辞下》中也提到“上古结绳而治，后世圣人易以书契，百官以治，万民以察”。

南美洲秘鲁古代用于计数的绳子叫作“克维普”。没有染色的“克维普”仅用于计数，染上色的“克维普”则表示一定的含义：黄色表示老玉米，红色表示武器等。

在古埃及，结绳计数还被用于制造直角。古埃及人在绳子上打 13 个结，得到 12 条线段，使得每段线段长度相等，利用勾股定理（勾三股四弦五）制造出一个直角。据说这个办法被应用于建筑上。

《鲁滨逊漂流记》中，鲁滨逊将木桩做成一个大十字架，树立在他第一次登陆的岸边。在这根方形木桩的四边，他每天用刀子刻一道槽痕，每 7 道槽痕中有一道比其余的长一倍，每个月第一天刻下的槽痕比那 7 天一道的长痕又长一倍。这就是鲁滨逊用刻痕计算时间的方法。

数觉或者计数都是对“数”本身的一种感觉和记录，并不能进行数之间的运算，而运算就涉及数的算术、进制和计算。

算术是一门“数（shù）”和“数（shǔ）数（shù）”的技术，比如在中国古代“数”表示计算用的竹（算）筹。算术包括的内容有：自然数的读法写法、进位制、记数法、自然数的基本运算，如分数与百分数计算、各种量及其计算、比和比例以及算术应用。

进制是规约的一种进位方法，目前我们熟悉的有二进制、八进制、十进制、十六进制等。事实上，可选用的进制很多，而且古人在设计各种进制的时候，或许并没有进行更多的理性思考，因为我们会发现古人发明的许多进制是与自然界的许多现象直接相关的，如表 1-1 所示。

表 1-1 进制与自然现象

序号	进制	自然现象
1	二进制	电平易分辨的高低两个状态
2	四进制	一年分四季
3	五进制	人的一只手有五个指头
4	八进制	二进制转换容易
5	十进制	人总共有十个手指
6	十二进制	十个手指 + 两只脚
7	十六进制	二进制转换容易
8	三十进制	月满月缺 28 + 天, 取整

据记载, 中国古代以十粒粟米并排的长度为一寸, 也就是说, 最小长度单位——分是以米粒的宽度为标准的。据推测, 重量单位可能也是以来来确定最小单位的, 以充填成年人一口的米粒重量计作一两 (临时性的重量单位)。看中国文字“两”的字形演变 (参见图 1-1, 备注: 字形演变示意图来自汉典网 www.zdic.net), 我们可以发现“两”字形是源于天秤的, 通过天秤, 我们可以知道, 有了 1 两, 就可以很容易得到另外 1 两, 2 个 1 两可以很快得到另外 1 个 2 两, 以此得到 8 两、16 两。按照这种倍增的做法, 人们发现如果 1 口的米粒重量为基本的一两单位的话, 大概 16 口米粒即可成为人一天的口粮, 这 16 口被定义为一斤, 所以在中国古代有“半斤八两”的说法, 或许就是源于这样的演义。

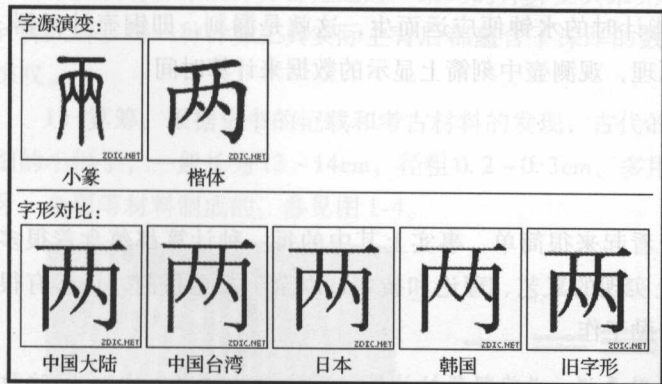


图 1-1 “两”字形演变及天秤

计算是一种将单一或复数输入值转换为单一或复数结果的思考过程, 包括自然的计算和算术的计算, 特别值得注意的是, 计算并不是算术专用的概念和名词, 甚至可以说, 古人对计算的理解都是先从自然的计算开始的。有趣的是, 古人对自然的计算一般都是从对时间的计算开始的。图 1-2 中的几个计时设备是完全利用自然界的特征设计的装置, 充分反映了古人的智慧。



图 1-2 圭表、日晷和铜壶计时器

1) 圭表是古代汉族科学家发明的度量日影长度的一种天文仪器，由“圭”和“表”两个部件组成。直立于平地上测日影的标杆和石柱，叫作表；正南正北方向平放的测定表影长度的刻板，叫作圭。当太阳照着表的时候，圭上出现了表的影子，根据影子的方向和长度就能读出时间。

2) 日晷本义是指太阳的影子。现代的“日晷”指的是古人利用日影测得时刻的一种计时仪器，又称“日规”。其原理就是利用太阳的投影方向来测定并划分时刻，通常由晷针和晷面组成。

3) 铜壶计时器：圭表和日晷都是用太阳的影子计算时间的，然而遇到了阴雨天或黑夜便失去作用了，于是一种白天黑夜都能计时的水钟便应运而生，这就是漏刻，即铜壶计时器。漏刻是以壶盛水，利用水均衡滴漏原理，观测壶中刻箭上显示的数据来计算时间。

1.3 模拟与数字计算

前面提到古人那种自然的计算，看起来很简单，事实上其中的每一种计算都蕴含着很多丰富的甚至至今我们无法在计算机上实现的工艺、理论和技术，以至于截至目前，仍然有很多的研究者孜孜不倦地在这个领域辛勤工作。

有一天，大发明家爱迪生对他的助手说：“劳驾帮忙计算一下这个电灯泡的体积。”这位助手是一个极具数学天赋的年轻人，接到这个任务后，这位助手拿出卡尺分段测量灯泡的各个横截面的周长，然后积分，进行了很长一段时间，仍然没有得出一个精准的结论。爱迪生看到后笑着说：“有这么复杂吗？把灯泡浸入一个满载水的容器中，然后测量一下溢出的水的体积不就可以了吗？”

另外一个例子，在化学分析领域中，有一个重要的仪器是色谱仪，其基本的功能是用电场和磁场将运动的离子按它们的质荷比分离后进行检测，测出离子准确质量，即可确定离子的化合物组成。基于这个原理，色谱分析法可以用于某样本的定性分析（有哪些成分）以及

定量分析（不同成分的比例），相对而言，定性分析比较困难，定量比较容易。参见图 1-3，每一种成分对应一个峰（其下的面积代表量），不同峰面积的比例表示不同成分的占比。有一天，老师 A 对学生 B 说：“你给我计算一下某两个成分的占比吧！”这位学生 B 拿到这幅色谱图后开始对各个峰面积进行拟合和积分（一种简单的方法是将这个色谱图印在一个等间距的方格图中，然后每个峰覆盖的方格数可以作为面积），折腾了很久也没有得到一个精确的结果。老师 A 看到后说：“这样吧，你去把每个峰用剪刀减下来，到天平上分别称重一下，质量比就是它们的面积比，不是吗？”

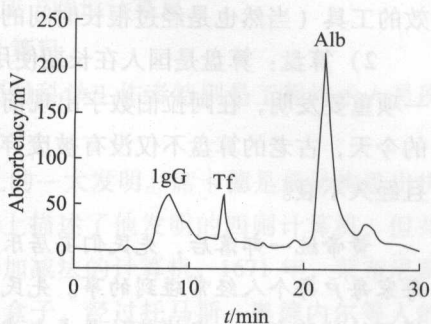


图 1-3 色谱图示意

上述的两个例子中，我们不能评论说爱迪生比助手聪明或者老师 A 比学生 B 聪明，因为他们采用的是两种不同的思维，爱迪生的做法和老师 A 的做法是一种典型的基于自然的计算，而助手和学生 B 是一种典型的基于算术的计算，两种思维方式没有优劣之分。事实上，现在的（高档）色谱仪已经能够在打印出色谱图的同时，将各个峰面积通过算术的方法积分出各个成分的配比。

所谓算术的计算是指通过算术而不是自然界的内生方法进行的计算。在人类发展的历史长河中，所有算术的计算都是通过一系列的计算工具来完成的，比如算筹、算盘、纳皮尔筹、计算尺等。每一种计算工具实际上背后都蕴含了深厚的数学基础，即使在今天看来仍然令人惊叹。

1) 算筹：根据史书的记载和考古材料的发现，古代的算筹实际上是一根根同样长短和粗细的小棍子，一般长为 13~14cm，径粗 0.2~0.3cm，多用竹子制成，也有用木头、兽骨、象牙、金属等材料制成的，参见图 1-4。

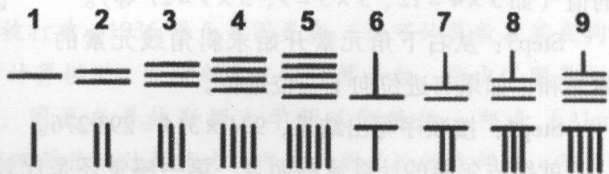
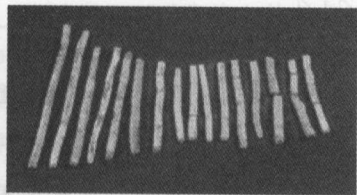


图 1-4 算筹及算筹计算示意

《孙子算经》记载，算筹记数法则是：凡算之法，先识其位，一纵十横，百立千僵，千十相望，万百相当。《夏阳侯算经》提及：满六以上，五在上方，六不积算，五不单张。

意思是说：算筹记数法是以纵横两种排列方式来表示单位数目的，其中 1~5 分别以纵横方式排列相应数目的算筹来表示，6~9 则以上面的算筹再加下面相应的算筹来表示。表示多位数时，个位用纵式，十位用横式，百位用纵式，千位用横式，以此类推，遇零则置空。这

种记数法遵循十进制。算筹的计算思想现在看来很简单，不过这在当时确实是一个非常有效的工具（当然也是经过很长时间的发展得到的）。

2) 算盘：算盘是国人在长期使用算筹的基础上发明的（公元前 600 年），是中国古代的一项重要发明，在阿拉伯数字出现前是广为使用的计算工具。即便是在计算机已被普遍使用的今天，古老的算盘不仅没有被废弃，反而因它的灵便、准确等优点，在许多国家广为使用且经久不衰。

黄帝统一部落后，先民们安居乐业，生产蒸蒸日上，物质越来越丰富，算账、管账成为每家每户每个人经常碰到的事。先民开始用石子计数（比如狩猎了多少羊或老虎等），但是其明显的缺陷在于：当捕获的羊或老虎过多的时候，管理难度变大，比如石子容易混乱，为了解决这个问题，第一个解决思路是用一根绳子把石头串起来（解决了石子混乱的问题）。这还不够，因为如果狩猎数目太多的话，一根绳子上则需要更多的石子。为了解决这个问题，又做了一次改良：每一根绳子上只有 10 个石子，更多的数目采用进位来实现。据说，这便是算盘最原始的形态。

随着时代不断前进，算盘不断得到改进，成为今天的“珠算”。特别是民间，即使认字不多，但是只要懂得了算盘的基本原理和操作规程，人人都会使用。

3) 纳皮尔筹：纳皮尔筹是 17 世纪英国数学家纳皮尔在著作里介绍的一种新工具，具体做法是：把格子乘法里填格子的工作事先做好，需要哪几个数字时，就将刻有这些数字的木条按格子乘法的形式拼合在一起。纳皮尔筹与中国的算筹在原理上大相径庭，它已经显露出对数计算方法的特征。

求解 934×314 。求解过程如下（参见图 1-5 左）：

Step1：首先将 934 和 314 分别横向和竖向放置。

Step2：在 3×3 的矩阵中依次计算每个元素的值（如 $3 \times 4 = 12$ 、 $3 \times 3 = 9$ 、 $3 \times 9 = 27$ 等）。

Step3：从右下角元素开始求斜角线元素的累加和，如果有进位向下一位进位。

Step4：按顺序写出结果， $934 \times 314 = 293\,276$ 。

就纳皮尔筹的计算流程而言，第一感觉还是比较巧妙的，而事实上对照一下算式 934×314 的标准计算过程（参见图 1-5 右），会发现两者的计算过程本质上是一致的。但是纳皮尔筹的这种处理方法，对后续的计算工具的影响，甚至机械的影响都极其巨大。

4) 计算尺：1614 年，对数被发明以后，乘除运算可以化为加减运算，计算尺就是依据这一特点来设计的模拟计算机，也叫对数计算尺，通常由三个互相锁定的有刻度的长条和一个滑动窗口（称为游标）组成。在 20 世纪 70 年代之前使用广泛，之后被电子计算器所取代，成为过时技术。几个标志性的计算尺如下：

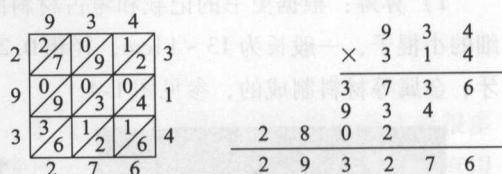


图 1-5 纳皮尔筹及普通乘法计算示意

- ①1620年, E. 冈特最先利用对数计算尺来计算乘除。
- ②1632年, 奥特雷德发明了有滑尺的计算尺, 并制成了圆形计算尺。
- ③1652年, R. 比萨克制成了有固定尺身和滑尺的计算尺。
- ④1850年, V. 曼南在计算尺上装上游标, 因此为当时科学工作者特别是工程技术人员所广泛采用。

5) 机械计算机: 与计算尺同时出现, 是计算工具上的一大发明。席卡德是最早构思出机械计算机的人(1623年), 他在给天文学家开普勒的信上描述了他发明的四则计算机, 但并没有成功制成; 1642年, 帕斯卡成功创制第一部能计算加减法的计算机; 1671年, 莱布尼茨发明了一种能作四则运算的手摇计算机, 是长1米的大盒子, 经过托马斯、奥德内尔等人的改良后, 多种多样的手摇计算机出现了, 风行全世界。17世纪末, 这种计算机传入了中国, 并由中国人制造了12位数的手摇计算机, 独创出一种算筹式手摇计算机。

上面简单介绍了历史上的几种典型计算工具, 可以发现计算工具的精巧度、工艺的复杂度和算法的蕴含度是随着历史的进程不断发展和进步的。当然, 所有的这些计算工具都是基于手工或基于机械的模拟计算工具。1946年, ENIAC (Electronic Numerical Integrator And Computer, 电子数值积分计算机) 的出现标志着电子计算机时代的到来。但在这之前, 堪称计算机界的教父——图灵做了一件伟大的事情, 从基本原理上提出通用的计算模型, 即图灵机。

数学家希尔伯特·西蒙 (Herbert Simon) 在1928年国际数学家大会上提出关于一阶逻辑公式可满足性的判定问题 (Entscheidungs Problem), 引起了很多数学家同行的关注。判定问题原型是: 希尔伯特计划把数学建立在一个完备的、一致的公理化系统之上。这意味着:

- 1) 所有的数学命题都能用符号无二义地表达出来。
- 2) 所有的数学命题都能被证明或证伪 (完备性)。
- 3) 对任意数学命题 P , 如果 P 被证明, 那么 $\neg P$ 必定能被证伪 (一致性)。
- 4) 如果最后再找到一个算法, 就能机械地判定一个数学命题的真伪 (可判定性)。

1931年, 数学家库尔特·哥德尔 (Kurt Gödel) 发表了震惊数学界的哥德尔不完备定理, 至此, 数学系统一致性和完备性的统一被打破。1936年5月图灵在《论可计算数及其在判定问题上的应用》中提出图灵机作为通用计算模型, 并重新定义可计算函数, 给出了图灵机无法判定的停机问题。1937~1938年间, 图灵在普林斯顿大学师从阿隆佐·邱奇 (Alonzo Church), 证明图灵机与邱奇的 λ 演算具有等价的计算能力。邱奇-图灵论题在邱奇、图灵与哥德尔三人的工作下成型, 该论题认为所有可有效计算的函数或算法都可由一台图灵机来执行。

邱奇-图灵论题表明图灵机能模拟所有机械的、有限步的计算。其对后世的影响在于:

- 1) 图灵机给计算机的具体实现提供了参考价值。
- 2) 算法问题从此有了坚实的基础。
- 3) 图灵机也给出了现在这个时代的计算上限。

阿塔纳索夫-贝瑞计算机 (Atanasoff-Berry Computer, 简称 ABC) 是法定的世界上第一台电子计算机 (原件损毁, 复制品永久展于艾奥瓦州立大学达勒姆计算和通信中心一楼大厅), 它是在 1937 年由艾奥瓦州立大学的约翰·文森特·阿塔纳索夫 (John Vincent Atanasoff) 和他的研究生克利福特·贝瑞 (Clifford Berry) 设计, 1942 年成功测试。该电子计算机的设计之初仅用于求解线性方程组, 不可编程, 示意图参见图 1-6。其基本特点包括:

- 1) 采用电能与电子元件 (电子真空管)。
- 2) 采用二进制制。
- 3) 采用电容器作为存储器。
- 4) 进行逻辑运算而非通常的数字算术。

可以看出, 阿塔纳索夫-贝瑞计算机包含了现代计算机中四个最重要的基本概念, 是一台真正现代意义上的电子计算机, 标志着人类计算从模拟时代挺进数字时代。

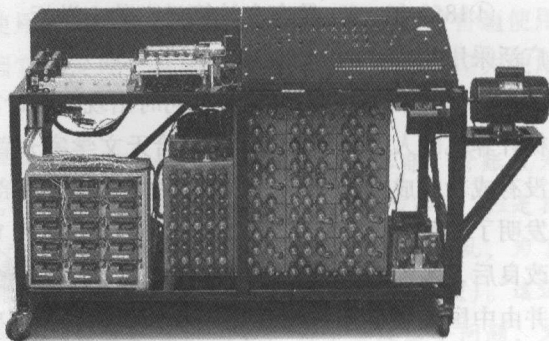


图 1-6 阿塔纳索夫-贝瑞计算机

二战期间, 美国宾夕法尼亚大学莫尔学院承担前线的火力表计算工作, 军方派戈尔斯坦 (Herman Heine Goldstine) 以军方代表身份驻扎在学校, 而事实上戈尔斯坦本人还是一位数学家。学校方负责火力表计算的是物理学院的莫奇利 (John William Mauchly), 他极其敏锐地感觉到电子计算机时代即将来临, 1942 年 8 月, 他在与戈尔斯坦的交流过程中提出了《高速电子管计算机装置的使用》备忘录。而戈尔斯坦也表示出了极大的兴趣, 于是在 1943 年 6 月 5 日, 莫尔学院与军械部正式签订合同, 该合同名就是“电子数值积分计算机 (简称 ENIAC)”。为了有效完成这个项目, 莫奇利邀约了电子学院的约翰·埃克特 (John Eckert) 和逻辑学家伯克斯 (A. W. Burks) 一起加盟项目组 (埃克特任这个项目的总工), 这四位组成的团队就是著名的莫尔小组。项目如期开展, 不过 ENIAC 存在两个问题: ①没有存储器; ②用布线接板进行控制大大降低了速度。一个偶然的机缘, 军方代表戈尔斯坦偶遇另外一个著名的数学家、经济学家冯·诺依曼 (John von Neumann) (据说他们是在一个火车站偶遇), 双方交流了一下彼此的工作近况, 冯·诺依曼对 ENIAC 产生了极大兴趣, 并在 1944 年受邀参加莫尔小组。1945 年, 他们在共同讨论的基础上, 发表了一个全新的报告《存储程序通用电子计算机方案 (Electronic Discrete Variable Automatic Computer, EDVAC)》, 这一报告的发表, 标志电子计算机时代的来临, 并且冯·诺依曼结构被作为实用的计算模型从此开始大行其道。

冯·诺依曼结构也称普林斯顿结构, 是一种将程序指令存储器和数据存储器合并在一起的存储器结构, 出自 1945 年 6 月冯·诺依曼的论文“First Draft of a Report on the EDVAC”, 其重要贡献集中体现在“程序存储+二进制”上, 具体而言, 该结构要求:

- 1) 必须有一个存储器。
- 2) 必须有一个控制器。

3) 必须有一个运算器,用于完成算术运算和逻辑运算。

4) 必须有输入设备和输出设备,用于进行人机通信。

1946年,第一台电子多用途计算机 ENIAC 在美国宾夕法尼亚大学诞生,ENIAC 是以电子管为主要电路元件的电子计算机,该发明标志着人类计算从模拟时代正式进入数字时代。

1.4 从数据到大数据

计算机的早期职能就是为军方服务,以数值计算为主。而随着时代的发展,单纯地为军方服务开始演变成为民用服务,单纯的数值计算也开始有了纯粹数值之外的数据类型的计算。

ENIAC 项目完结后,由于价值观的不同,莫尔小组的四位成员和冯·诺依曼分成两个分支。冯·诺依曼、戈尔斯坦和伯克斯继续围绕 ENIAC 进行更具理论的研究,他们在为普林斯顿大学高级研究所研制 IAS 计算机时,又提出了一个更加完善的设计报告《电子计算机逻辑设计初探》。而埃克特和莫奇利在 1947 年联合成立了“埃克特-莫奇利计算机公司”(Eckert-Mauchly Computer Corporation)(或许是最早的计算机公司),专门从事 ENIAC 后续的改进和产业化,并将其命名为商用 UNIVAC,承担了美国人口统计局的人口统计项目。这个项目的意义至少在于:计算机的应用开始向非军事用途蔓延。与埃克特-莫奇利计算机公司同时竞标此项目的另外一个伟大公司(IBM)也敏锐地发现这个时代已经进入了计算机的时代,主营开始向计算机转型。当然,这个项目以及“埃克特-莫奇利计算机公司”的运作并不理想,这就演绎出后续的许多桥段,此处不再赘述。

所有录入计算机的数据首先必须数字化。所谓数字化是指将模拟信号转化为数字信号的过程,包括采样和 A/D 转化(模拟到数字转化)两个过程。计算机在进行数学运算时采用的是二进制,二进制的所有数都用字符 0 和 1 的组合表示,基本单位是位(bit)。如何进行采样和如何进行 A/D 转化不是本文涉及的内容,故在此不赘述。不过需要说明的是:此处的模拟信号指的是电信号,至于普通物理信号如何变为电信号则是传感器和变送器的话题。

在数字化的基础上便是数据化,所谓数据化是将“位”结构化和颗粒化,形成标准化的、开放的、非线性的、通用的数据,其基本单位是字节(byte)。前面已经介绍过,数据是指存储在某种介质上能够识别的物理符号(数、字符或者其他),是信息的载体。

随着科技的进步,尤其是材料学、物理学、化学以及信息学等的不断发展,数字化的手段越来越多。更重要的是,随着摩尔定律的持续发酵,数字化的成本越来越低,最终使得在可以接受的成本控制条件下,万事万物皆可数字化,并且是基于不同的原理、从不同的角度进行数字化,这直接导致数据的类型越来越多、数据获得的渠道越来越多、数据采集的速度越来越快。

另一方面,人类的需求也处于不断膨胀中,人们希望有更多的渠道了解自己、了解自己的朋友、了解自己生存的环境,人们也希望有更多的手段和其他人、其他环境交互,人们希

望能够更舒适、更和谐地工作、学习、生活、商务、社交。总之，我们会发现，随着科技的进步，加之需求的膨胀，人们发明创造了很多的仪器、设备甚至软件平台，而这些仪器、设备及软件平台每天都在产生大量的数据，每个人每天在生产，同时也在享受和消费各种类型的数据，每天都我们都生活在数据的海洋中。

每天我们通过电话或短信与朋友交流，我们在使用电信公司通信服务的同时，通话痕迹或者短信痕迹就保留在电信运营商的数据库里。

每天我们通过社交网站发布自己的状态，当然也能看到朋友的状态，我们在通过社交平台与朋友交流的同时，这些状态或互动数据保留在社交平台的后台数据库里。

或许我们会去医院，医院各个检测设备可以对我们的体征进行各个维度的检测，医生依据这些数据对我们的健康状况进行评估，我们获得医疗服务的同时，健康数据保留在医院的数据库里。

我们在购物网站上购物，在享受便捷购物的同时，我们的浏览记录、支付记录保留在了购物网站的后台数据库里。

或许我们会用信用卡在实体店购买物品或者去餐馆吃饭，而信用卡支付记录也就保留在了各个银行的数据库里。

即便不使用信用卡而仅仅是使用现金，那么每次到银行做各种交割业务，我们办理的所有的业务记录信息就这样保留在了银行的数据库里。

即便我们仅仅是去逛逛街、逛逛商场，各个安防监控的视频监控系统也将我们的影像记录在后台数据库中。

我们或许还要远行，反映我们行程轨迹的机票订单、火车票订单、旅馆订单等均在我们享受各类服务的同时被保留了下来。

如使用导航，那么我们的出行轨迹甚至出行偏好也就保留在导航服务公司的后台数据库中。

如果我们仅仅从家去单位，沿途的各类监控也把我们的影像记录下来并保留在了后台数据库。

如果我们带着手机，手机需要和不同的基站切换以便让我们一直能够得到通信服务，而同时，我们的行动轨迹就记录在了基站里。

我们每天总是会使用水、电、煤等，我们每月的消费量以及我们的支付方式就被记录和保留了下来，甚至能够保留每个月、每一天、每一小时消费数据。

如果我们有车，我们的车辆归属信息就被保留在了公安交管信息系统中。

如果我们买房，我们的房产信息就被保留在了房产局系统中。

任何一个人还有户口，我们的户籍信息，家中有几口人、几个小孩、几个老人等就保留在了公安户籍系统中。

即便我们睡着了，可能戴着的手环也会把我们的睡眠数据记录下来……

我们每天在生产数据的同时也留下了反映我们衣食住行商旅乐的各种痕迹，这是我们无法回避的事实，因为我们需要享受这些服务，而在享受服务的同时，痕迹总是在你毫无感知的情況下悄悄地被保留下来了。自然人是这样，企业法人也是这样。

企业的成立需要注册，那么企业的相关信息就被保留在了工商系统数据库中；企业总要缴纳税务，企业的财税信息就被保留在了税务机关业务系统中；企业总要做生意，企业的往来账目就在银行交割的同时被记录了下来；当然企业也要用水和电，有的还要用煤和气，这样企业的能源消耗数据就被保留进了相应的能源公司系统里；企业总归要聘用雇员，雇员的工资、税务、社保、保险等，所有这些数据就被记录在相应的业务系统中；或许企业也要做形象宣传，那么企业主动推送的各类信息就被保留在了网站门户、微博等中；或许公司还会去打官司，那么关于企业的官司信息至少会被存放在公、检、法各级业务系统中……

企业法人无法回避的事实是：公司经营、运维的整个过程均在生产数据的同时也留下了反映公司经营行为、信用等各种痕迹的数据。企业是这样，我们的政府乃至所有国家的政府都是这样。

普通的百姓群众无法参加各级常委的各类会议，也就无从知晓每次会议的主题和精神，但总是有工作人员记录并作为会议纪要留存，其实也是被记录下来，只是普通百姓无法获得；政府总是要做事的，那么政府的建议、决定、指南以及各种交流活动总会主动地推送到门户网站上或者被各级记者采访记录在各大媒体中；政府为了完成其固有的职能总有一整套完整的组织结构、人员架构等，而仅仅从管理的业务来看，其工资、税务、社保等信息也被保留在各个信息系统的后台数据库中；各级政府的执政业绩总归是要反映到每年的财报中，而这种财报总是以某种形式记录，不论其是以纸质版本记录在档案里还是记录在某个信息系统数据库中，又或者记录在普通百姓的心中……

因此，无论是政府、企业法人还是自然人都回避不了的事实是：各个利益角色都被数字化和数据化了。

所有的数据都是有价值的（价值本身具有较强的主观性，不同的利益角色出于不同的需求会表现出不同的价值期望），而且这种价值不仅仅是该数据产生时的价值。比如在微信朋友圈或者微博上发布的状态，原本的用途只是告诉朋友圈的朋友们当时周边的风景、感悟、心得等，但数据更大的价值在于通过对更多数据的采集和整合，可以实现对单个的自然人、企业法人等进行全方位的刻画和描述（有的场合称为“360度画像”），借此实现对自然人、企业法人的价值度、信用度等评估，实现更多的商业用途。更多的关于数据价值的内容参见11.2节。

我们在购物网站上购买了某个物品，甚至我们仅仅在搜索引擎中搜索了某类产品的关键字，我们就有可能在打开网页的时候被推送一系列的相关产品广告，广告的发布者就是利用了我们的购买行为和检索行为探测出我们对某类产品有兴趣，或者预测出我们购买了某些产

品后还应该购买其他的一些产品。比如,曾经购买过鱼竿,与鱼竿相关的钓鱼线、鱼饵的产品或许也是潜在的购买对象,于是就可以推送诸如此类的产品广告。更甚的是,如果商家通过其他渠道获知与我们类似的人都在购买某个产品,商家或许就会将类似的产品广告推送给我们,想法其实也很简单:相似的人应该有相似的产品偏好和消费习惯……

上述的例子是一个典型的自动推荐和精准营销的案例。为达成此目标,商家需要对产品聚类,并且对消费人群进行分类,借此实现将合适的产品推送给合适的人的商业目标。上面的这个例子可能还不足够完备,因为每个人的消费能力、消费偏好可能完全不一样,每个人对广告推送的渠道也有不同的偏好,每个人在不同时机对接收到广告接受度也不一样,因此利用合适的渠道、在合适的时机、向潜在目标人群精准推送产品的广告信息或许会带来更大的商业利益。而这种更为精准的自动推荐则需要有更多的数据来更具体、客观、完备地描述人、物、渠道等。因此,从商家的经营视角而言,需要更多的、更能对目标进行刻画和描述的数据。

上面的例子还表明,为了某个具体的商业目标,需要不同数据源的不同数据。而不同数据源数据的组合也能够满足不同的商业目标。仍以在电商网站上购物为例,我们把不同地域的人群购买的物品进行聚类就可以获悉对应地区的消费偏好,这对于厂家和商家的物流配送中心的布局就有极大的帮助。

“啤酒-尿不湿”的故事是数据挖掘中的一个经典案例,故事的原型是这样的:商家通过超市的售货记录突然发现(技术实际上很简单,就是经典的关联规则挖掘),在所有消费群体的购买行为中,买尿不湿的往往同时还会买啤酒。两种本来风马牛不相及的产品为什么会出现在一个销售小票中呢?不管原因是什么,商家总可以利用此规律对商品的配比及摆放做出一些有价值的决策。

这个故事还在发展——商家又发现另一个有趣的规律:每当飓风天气来临的时候,商家的蛋挞卖得很快,而且通过销售记录分析,发现买蛋挞的同时购买手电筒的比例也很高。这个故事被称为“飓风-蛋挞”而广为流传。分析其中的原因或许也很简单:飓风天气来临,交通不便,于是我要多储备些吃的;飓风天气可能会导致供电故障,所以要购买手电筒备用。

上述这两个例子看似都是在根据顾客的购买行为分析出了“啤酒-尿不湿”的关联和“飓风-蛋挞”的关联或“蛋挞-手电筒”的关联,其本质区别在于:“啤酒-尿不湿”的分析数据源仅限于超市的售货记录(这是超市的内部数据);而第二个例子除了利用超市的内部数据外,还利用了气象数据,对于超市而言,这个数据源是典型的外部数据。通过这个例子,我们可以看出,更多的数据源(内部数据+外部数据)能够挖掘分析出更多的、有价值的或者是仅仅有趣的结果。

综上所述,人类文明进程的不断发展和人类需求的不断膨胀(在数据层次需要更多的数据支撑),万事万物数字化、数据化(数字化手段越来越多、数据获取渠道越来越多、数字化成本越来越低),数据交叉复用化(可以赢得更多商业利益),这些直接引导我们进入了现在所称的“大数据时代”。

1.5 大数据时代

马云说：“大家还没搞清 PC 时代的时候，移动互联网来了，还没搞清移动互联网的时候，大数据时代来了。”

事实上催生出大数据时代的来临还有几个重要的事实：

1) 1981 年 8 月 12 日，IBM 公司正式推出了全球第一台个人计算机 IBM PC，这直接使得计算机从很神圣的高端计算走进民间，从纯粹的计算走向管理和娱乐。IBM PC 的诞生还直接催生出两大 IT 巨头：Intel 和微软。

2) 1990 年 12 月 25 日，蒂姆·伯纳斯·李（Tim Berners-Lee）在日内瓦的欧洲粒子物理实验室用 HTML 及 HTTP 开发出了世界上第一个网页浏览器，万维网的普及推动了互联网向各个领域的普及和渗透（关于互联网相关详细信息参见第 13 章）。

3) 随着互联网技术的发展及互联网的持续普及，移动通信技术的快速发展将计算机历史从互联网时代拉入了移动互联网时代。

可以说移动互联网时代的来临直接推进了大数据时代的到来。一方面移动互联网本身就催生出了大数据，而大数据也持续推进和优化移动互联网的不断发展。更为重要的是，移动互联网的到来使得用户衣食住行商旅乐的习惯和方式均产生了巨大的变化。因此，必须有匹配的策略、手段和技术响应这些特征的变化，比如创新的应用模式、商业模式和运营模式等；而这些需求也“倒逼”政府、工业界和学术界对这一时代特征进行有针对性的、有建设性的、审慎务实的策略制定、技术预研和产品规划。

酒后代驾业务模式事实上历史并不长，但在不长的历史中，我们就已经看到其模式从传统的熟人介绍到后来的电话预约，再到现在的手机终端 APP 预约。我们无法预测未来会有如何的变迁，但是应用模式的每一次变迁都使得用户在享受“酒后代驾”服务方面变得更便捷、更高效，而作为“酒后代驾”的服务方，比如司机也因为能够更便捷和高效地获知周边的需求而获得更高收益。

商业模式也在不断变迁。传统上我们使用某一个软件一定要去购买，而如今，我们使用很多的应用（尤其是手机终端 APP）大多是免费的，那么商家的利益谁来保障？事实上我们无须认为这些商家在做慈善，商家的目标都是要获益，只是因为时代的变迁，商业模式也在变迁而已，比如“免费模式”“后端收费模式”“羊毛出在猪身上”等。

纯技术的开发也是这样。以软件开发为例，为了响应“软件危机”，“软件工程”作为一门学科被提出和发展；为了响应需求的演化，“敏捷软件工程”被提出和发展；而进入移动互联网阶段，手机终端 APP 的开发模式仅仅使用“敏捷软件工程”开发模式也不足够，也就引出“互联网开发模式”这样的话题。

从计算机科学与技术这门学科的角度出发，我们可以看到，从天才的图灵发明伟大的图

灵机开始, 计算机科学与技术的研究一直按照相对清晰的思路在发展, 即硬件和软件: 一方面从硬件的角度出发衍生出计算机组成原理、计算机体系结构等; 另一方面, 从软件的角度出发衍生出软件方法学、程序设计、数据结构等。根据应用目标的不同也衍生出了不同计算机应用技术学科, 如人工智能、数据库、自然语言理解等。

随着时代的发展和变迁, 我们突然发现从计算机这一个中心出发衍生的内容太多了, 以至于我们无法归纳出像以前那样很清晰的定义, 比如什么是计算、如何计算等。于是我们会发现很多的技术名词如过眼云烟一样也在不断地变迁, 比如: 分布式计算、网格计算、弹性计算、透明计算、普适计算、无所不在计算、服务计算、云计算……当然每一种概念的提出都是针对某一个具体的问题场景或技术挑战进行的, 并且每一种概念的提出都是先前概念技术的发展和延伸, 并且有自己独立的边界。问题是: 随着时代的发展, 我们面临的需求和挑战不断被扁平化、个性化和垂直化, 按照目前的技术脉络逻辑, 我们需要更多的“X 计算”。那么问题来了: $X = ?$

大数据的来临意味着我们科学研究所依赖的范式需要改变, 而这种范式的改变或许会是响应上述问题的一个契机。数据科学就是在这一背景下诞生的, “数据科学”这一概念的出现凸显了各界对大数据的关注和期待。

图灵奖获得者吉姆·格雷 (Jim Gray) 在总结了科学研究在人类历史上所先后经历的实验科学、理论科学和计算科学三个范式之后, 提出了基于数据和思维的第四范式: 数据密集型科学发现 (Data-Intensive Scientific Discovery); EMC 公司认为数据科学必将成为大数据分析的持续驱动力; 李国杰院士对数据科学的描述是, “计算机科学是关于算法的科学, 数据科学是关于数据的科学”。

关于第四范式及数据科学的话题, 在 9.3 节中有详细介绍, 此处不再赘述。总之, 大数据时代来了。

大数据是这个时代的特征, 各界都在热议大数据, 而事实上, 作为一种现象或者一个名词, 大数据并不是一个新鲜的事物。如前所述, “大数据是大到无法通过人工在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息”, 按照这样的朴素描述, 历史上出现过很多次类似的挑战和现象。

在人类走向文明的历程中, 那时还没有规范化的语言, 人与人之间的沟通均通过手势、声音来进行, 很显然, 对于部落、小范围的沟通而言, 这是容易交流、便于协作的, 但是明显的缺陷在于: 地理范围、对象范围有限, 容易失真、失传。这个时候如果希望在更大范围内进行交流和沟通时, 当时的人类就遇到了“无法利用传统的手段实时有效地完成交流”的问题, 我们暂且认为这个问题是那个时代的“大数据”。还好因为人类的智慧, 发明了“语言”, 解决了上述问题; 也因为语言的诞生, 人类从此走进了文明。

在人类走进文明的历程中, 那时已经有了规范的语言, 人与人之间的沟通可以通过语言来进行, 于是人类的意图、想法、思想可以更广泛、更精准地传播和传承, 但是不可避免

问题是：记录成本昂贵，难以全民推广。当时的人类就遇到了“无法利用传统的手段实时有效地完成传承”的问题，我们暂且认为这个问题是那个时代的“大数据”。还好因为人类的智慧，发明了“印刷术”。

印刷术结束了手稿时代，让文化广为传播，再次扩充了信息的数量和组织的规模，使得人类的文明成果以更快、更保险的方式传播，但是存在的典型问题是：数据组织复杂、数据泛滥严重，或许这也算是当时无法解决的一个“大数据”问题。好在信息技术的引入，让书籍的管理变得有序、规范和便捷。当然信息化时代的到来，也引发很多的挑战和困难，这就是后话了。

1980年，阿尔文·托夫勒（Alvin Toffler）在《第三次浪潮》中指出，“如果说IBM的主机拉开了信息化革命的大幕，那么大数据则是第三次浪潮的华彩乐章。”这或许是“大数据”作为专业术语第一次被单独提出。

1997年，迈克尔（Michael）等人指出可视化领域中数据集通常相当大，占据了主内存、本地磁盘甚至远程磁盘的容量，他们将此问题称为“大数据”，这是ACM数据图书馆的第一篇使用“大数据”术语的文章。

1998年，SGI的首席科学家约翰·马西（John R. Masey）在USEUIX会议上指出数据存储的高速增长及即时访问需求的迅速增长将对物理和计算的基础设施产生巨大压力，同时也将为研究和商业应用带来巨大机遇。

2000年，弗朗西斯（Francis）等人指出很多科学如物理、生物、社会学等从“大数据”中获益，并指出大数据是指数量（或质量）在快速膨胀的可用、潜在的相关数据。

2001年，道格·兰尼（Doug Laney）发表了“3D Data Management: Controlling Data Volume, Velocity, and Variety”研究报告，其“3V”已成为普遍接受的关于大数据的三个定义特征。

2008年《Nature》出版专刊“Big Data”，从互联网技术、网络经济学、超级计算、环境科学、生物医药等多个方面介绍了海量数据带来的挑战。

上面简单罗列了大数据作为一个名词在不长的历史片段中的一些关键节点，或许还不够完备，但已经可以看出，在大数据这件事情上，工业界、学术界一直在关注，后文的相应章节也会专门介绍。

前文给出了大数据的一般理解是：大数据是大到无法通过人工在合理时间内截取、管理、处理并整理成为人类所能解读的信息。关于大数据的定义事实上并不唯一，不同的角色从不同角度对大数据都有不同的理解，比如：

1) “Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.”, *Oxford English Dictionary* (OED).

2) “An all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.”, *Wikipedia*, 2014.

3) “Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”, *McKinsey*.

4) “The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value” and “...things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value.”, *Viktor Mayer-Schönberger and Kenneth Cukier*.

5) “The broad range of new and massive data types that have appeared over the last decade or so.”, *Davenport*.

6) The belief that the more data you have the more insights and answers will rise automatically from the pool of ones and zeros.

7) A new attitude by businesses, non-profits, government agencies, and individuals that combining data from multiple sources could lead to better decisions.

8) The new tools helping us find relevant data and analyze its implications.

9) The convergence of enterprise and consumer IT.

10) The shift (for enterprises) from processing internal data to mining external data.

11) The shift (for individuals) from consuming data to creating data.

12) The merger of Madame Olympe Maxime and Lieutenant Commander Data.

从上述 12 个关于大数据的理解和定义（其中定义 6~12 来源于 <http://www.forbes.com/sites/gilpress/>）来看，所有定义都是从以下三个角度进行描述的：

1) 大数据是什么？这是对大数据本质属性的一个描述，无论是相对还是绝对，大数据的量级都很大、种类都很多，比如定义 1~5。

2) 大数据有用吗？这是对大数据目标属性的一个描述，大数据一定是有用的，数据交叉复用一定会带来价值，这是一种信念、一种期待，比如定义 6~7。

3) 大数据如何用？这是对大数据技术属性的一个描述，大数据是一种采集、整合、分析数据的手段，比如定义 4、8~10。

有一种说法，相对比较完备地覆盖上述惯有大数据的理解和定义。这种说法是：所谓大数据，就是指数据本身及为了实现“数据→价值”这一价值逻辑而涉及的工具、平台和系统的合集。在大数据最原始的定义中提及的 4V 特征，也很好理解：

1) 随着信息技术的发展，数字化手段越来越多，不同的数字化设备以不同的方式记录和描述目标对象，这就引发了数据类型越来越多的事实，这对应于大数据 4V 特征中的 Variety。

2) 随着信息技术的发展，数字化成本越来越低，采集性能越来越好，体现在采集的速度及实时性方面，短时间内可以采集更多的数据，这对应于大数据 4V 特征中的 Velocity。

3) 信息技术的不断发展推动了人类需求的不断膨胀，人们希望从更多的渠道采集数据，也就是说数字化需求越来越多的事实引发了数据规模越来越大的可能，这对应于大数据 4V 特征中的 Volume。

4) 人类需求的不断膨胀,体现在数据层的交叉复用的需求逐步变大以及从数据中获得价值的期望也越来越大,这对应于大数据4V特征中的Value。

综上所述,具有不同价值观、知识观、应用背景、技术背景及思维方式的不同角色,对于大数据的理解都有属于自己角色特点的大数据定义,换句话说,每个人心中都有一个关于大数据的定义,不是吗?

1.6 本章小结

如图1-7所示,人类利用智慧,从最原始的数觉,发明了计数的方法,然后发明了算术、模拟计算和数字计算。从另外一个维度上来看,因为需求的膨胀,人类从最原始的对数值的关注到对数据的关注,以及到目前对大数据的关注,一个典型的理念变革在于:在数据时代,计算是中心;而在大数据时代,数据是中心。所有这些理念的变迁或许本质上是人类需求不断丰富、软硬件技术不断发展的共同结果。

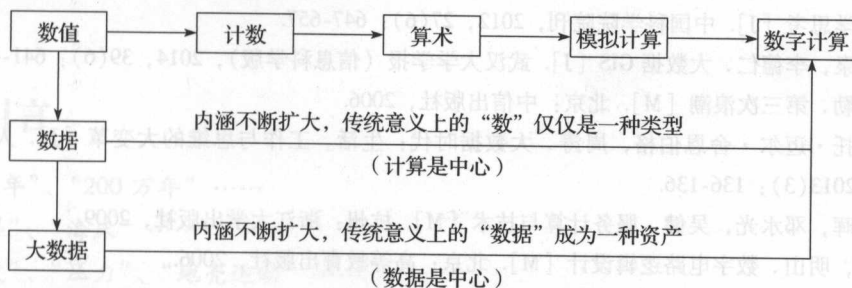


图1-7 大数据的发展历程

如上所述,数据是指存储在某种介质上能够识别的物理符号(数、字符或者其他),是信息的载体。而大数据是大到无法通过人工在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息。那么一个简单的问题是:大数据=“大”的数据?或许不是,否则不会在大数据的定义之后一定要加上所谓的4V特征(或者更多);或许也是,因为这里的“大”不是简单的“big”,至少就中文“大”而言,它有着丰富的哲学内涵。

《道德经》记载:“大白若辱,大方无隅,大器晚成,大音希声,大象无形。”这是老子提出的中国古代文学理论中的一种美学观念,意在推崇自然的而非人为的美。但显然,其中关于“大”的描述反映了中国人的智慧,值得品味。

众说纷纭,每个人都有自己的大数据定义。总之,大数据时代来了!

本章参考文献

- [1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 284(5): 28-37.

- [2] Campbell-Kelly M, Aspray W, Ensmenger N, et al. Computer [M]. Boulder, Colorado: Westview Press, 2013.
- [3] Gustafson J. Reconstruction of the Atanasoff-Berry Computer [J]. The First Computers: History and Architectures, 2000: 91-106.
- [4] Hadnagy C. Social Engineering: The Art of Human Hacking [M]. Manhattan: John Wiley & Sons, 2010.
- [5] Nonaka I, Toyama R, Konno N. SECI, Ba and Leadership: A Unified Model of Dynamic Knowledge Creation [J]. Long Range Planning, 2000, 33(1): 5-34.
- [6] Scardamalia M, Bereiter C. Computer Support for Knowledge-building Communities [J]. The Journal of the Learning Sciences, 1994, 3(3): 265-283.
- [7] Von Neumann J. First Draft of a Report on the EDVAC [J]. IEEE Annals of the History of Computing, 1993(4): 27-75.
- [8] 韩家炜, 坎伯, 裴健, 等. 数据挖掘: 概念与技术 [M]. 北京: 机械工业出版社, 2007.
- [9] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考 [J]. 中国科学院院刊, 2012, 27(6): 647-657.
- [10] 李清泉, 李德仁. 大数据 GIS [J]. 武汉大学学报 (信息科学版), 2014, 39(6): 641-644.
- [11] 托夫勒. 第三次浪潮 [M]. 北京: 中信出版社, 2006.
- [12] 维克托·迈尔·舍恩伯格, 周涛. 大数据时代: 生活、工作与思维的大变革 [J]. 人力资源管理, 2013(3): 136-136.
- [13] 吴朝晖, 邓水光, 吴健. 服务计算与技术 [M]. 杭州: 浙江大学出版社, 2009.
- [14] 毓银, 明山. 数字电路逻辑设计 [M]. 北京: 高等教育出版社, 2006.

大数据现象

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系张建兵老师及智能信息处理研究组的王茜、汤兆亮、王强、冯艺琳等几位同学的协助，在此表示深深的谢意。

2.1 引言

“5 亿年”、“200 万年”……

“海洋”、“海床”……

“温度”、“压力”、“地壳运动”……

“古生物”、“遗体”……

上述的一系列名词看起来风马牛不相及，但却正是石油形成的关键词。

研究表明，石油的生成至少需要 200 万年的时间，在现今已发现的油藏中，时间最老的达 5 亿年之久。大多数地质学家认为，石油像煤和天然气一样，是古代有机物通过漫长的压缩和加热后逐渐形成的。按照这个理论，石油是由史前的海洋动物和藻类尸体变化形成的，经过漫长的地质年代，这些有机物与淤泥混合，被埋在厚厚的沉积岩下。在地下的高温（备注：这个高温是在一个称为“油窗”的温度范围条件下。太低，石油无法形成；太高则形成天然气）和高压下它们逐渐转化，首先形成蜡状的油页岩，后来退化成液态和气态的碳氢化合物。由于这些碳氢化合物比附近的岩石轻，它们向上渗透到附近的岩层中，直到渗透到上面紧密无法渗透的、本身则多空的岩层中。这样聚集到一起的石油形成油田。

上述生物成油理论也被称为“罗蒙诺索夫假说”。这个假说是否合理（比如有反对的声音认为，即使把地球上所有的生物都转化为石油，成油量与地球上探明的储量也相差过大），本文不加关心。我们看到的是石油成了人类最重要的能源资源，世界各国因为石油而进行的战争、角力、博弈不计其数。第四次中东战争、两伊战争、伊拉克战争无不是如此，可以说石

油与国际政治息息相关。

在大数据时代，出于对大数据的重视和战略解读，将大数据作为一个“未来的新石油”的战略资源已经成为各国政府的共识。

一家电信运营商使用社交分析方法，筛选了3.65亿个电话记录，找出可能流失的客户并提供针对性的服务，大大提高了季度收益；一家金融服务公司从570亿笔ATM交易中检测出了欺诈模式……这是体现大数据价值的一些典型例子。数据已经成为一种新的经济资产类别，就像货币或黄金一样，将形成数据材料、数据探矿、数据加工、数据服务等一系列新兴产业。

在经历了几年的批判、质疑、研讨和炒作之后，大数据终于迎来了属于它的时代。2012年3月22日，奥巴马政府宣布投资2亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家战略。奥巴马政府甚至将大数据定义为“未来的新石油”。

2012年美国总统选举，投票前，这次总统选举一直被认为因为选情太接近而无法预计哪方会获胜，许多评论员强调无论是奥巴马还是罗姆尼都有获胜的可能。然而，纳特·西尔弗(Nate Silver)不这样认为。经过持续几个月来的预测，在投票当天，他预测奥巴马将有90.9%的概率获得大多数选票，如果按州计算，他准确预测了所有州的选举结果。本次大选预测的成功并不仅是纳特·西尔弗个人的胜利，更是大数据时代来临的标志。纳特·西尔弗凭着他的大数据模型，单枪匹马打败了整个政治阶级——时政记者、政党媒体顾问、雇佣文人以及政治评论员等。

2012年奥巴马连任总统后，立刻发布了“大数据研究和发展倡议”(Big Data Research and Development Initiative)，不知道这个倡议计划是否受到了纳特·西尔弗大数据模型的启示？当然，这是另外的桥段，此处不再赘述。大数据时代的另外一个脍炙人口的例子是Google的冬季流感预测，这个故事大致如下：

2008年，Google通过分析5000万条美国人最频繁检索的词汇，将之与美国疾病中心在2003年到2008年间季节性流感传播时期的数据进行比较，并建立一个特定的数学模型。最终Google成功预测了2009年冬季流感的传播，甚至可以具体到特定的地区和州，根据疾病预防控制中心的事后评估，其精准度高达97%。这个研究成果发表在2009年2月的《自然》杂志上。

这个桥段被认为是大数据的一个重要应用场景，也是类似这样的故事引起了工业界和学术界对大数据更多的关注和重视。Google成功预测冬季流感的一个基本思路是：人们输入的搜索关键词代表了他们的即时需要，反映出用户情况。为便于建立关联，设计人员编入“一揽子”流感关键词，包括温度计、流感症状、肌肉疼痛、胸闷等。只要用户输入这些关键词，系统就会展开跟踪分析，创建地区流感图表和流感地图。

当然，在后续的流感预测中，Google的方法失灵了，比如2013流感预警严重出错。人们不禁怀疑：一直热捧的“大数据”怎么如此不堪？经过理性研判，可以知道，有很多原因导致了2013流感预警严重出错，而其中的一个重要原因或许是：搜索引擎的开发者为了便于用

户的使用,而在用户实时输入检索词的过程中,给出一些推荐的检索词。由于这种推荐是精准的,意味着用户很大可能会选择系统推荐的关键词进行搜索。而从流感预测这个目标来看,其所依赖的反映用户即时需求的关键词数据事实上被搜索引擎本身加工过,实际上已经不是反映用户真实需求的数据,因此,预警出错也不足为奇了。这个桥段对我们的提示可能在于:①我们要尽可能收集“真实”的数据;②我们要尽可能收集“原始”的数据;③我们不要人为干预数据产生的过程。

大数据时代另一个脍炙人口的例子是塔吉特公司的精准营销。

明尼苏达州一家塔吉特门店被客户投诉,一位中年男子指控塔吉特将婴儿产品优惠券寄给他的女儿——一个高中生。但没多久他却来电道歉,因为女儿经他逼问后坦承自己真的怀孕了。塔吉特百货就是靠着分析用户所有的购物数据,然后通过相关关系分析得出事情的真实状况。

一个朝夕相处的父亲都没有及时了解女儿的状况,而作为局外人的塔吉特百货仅通过数据的相关性分析,却成功预测并进行了精准营销。这个桥段表明:大数据及大数据分析本身是有价值的。通过数据关联分析及建模,能够对人进行更全面和立体的刻画,为精准广告营销提供数据支撑。事实上,广告营销也被认为是大数据落地应用的重要场景之一,这在后文会有详细叙述。

2014年,SAP研发的Match Insight助力德国足球队夺得大力神杯是另一段大数据佳话:

2013年,德国足协与SAP开启合作之旅。当时,德国队领队奥利弗·比埃尔霍夫(Oliver Bierhoff)就在球员更衣室里展开了一番“市场调查”。他在调查中发现球员更喜欢通过数字平台沟通。因为沟通是比赛准备阶段最重要的环节之一,于是比埃尔霍夫委托SAP开发一款应用,旨在帮助球队沟通日程信息以及了解竞争对手的数据。SAP成功开发了一款基于SAP HANA的SAP Match Insights应用,该应用能够同步足球播报员播报的数据与球场视频片段所捕获的数据。

在短短6个星期内,SAP进一步开发了SAP Match Insights的新功能,使教练、工作人员和球员均能有效利用数据。球员和教练很快便对SAP Match Insights爱不释手。德国足协还在训练中心的休息室架起了大尺寸的触摸屏,以便球员和教练操作。最重要的是,他们还能将应用下载到移动设备上,这样球队就能随时随地获取数据。借助SAP Match Insights和SAP HANA,德国队教练简化了球队训练,进而提升了球队表现。这不仅让精彩的世界杯更加魅力四射,也为全世界的球迷献上了一场顶级视觉盛宴。

伴随着大数据时代的来临,世界各国对数据的重视提升到了前所未有的高度。套上大数据的光环后,原本那些存放在服务器上平淡无奇的“陈年旧数”一夜之间身价倍增。数据是数字时代的“石油”和“黄金”,对未来社会发展起着至关重要的作用:

1) 手中握有数据的公司站在金矿上,基于数据交易即可产生很好的效益。

2) 基于数据挖掘的各种商业模式争相诞生,大有百家争鸣的景象。

3) 更重要的是大数据对社会经济生活产生的影响绝不限于技术层面, 它为我们看待世界提供了一种全新的方法。未来, 数据可能成为最大的交易商品, 数据将成为一切行业当中决定胜负的重要因素, 数据已然成为人类至关重要的战略资源。

政府出台一系列出于国家意志的大数据战略、方针、政策, 为什么? 业界为何对大数据情有独钟? 百家争鸣的学术界也对大数据表示出了极大的热情, 甚至提出了诸如第四范式及“数据科学”这样的新兴概念, 为什么? 大数据得到各界如此高度关注的缘由在哪里? 本章尝试罗列大数据这一概念得到热炒的当下, 因为不同的利益驱动使然而引发“政产学研商用”各界迎接和拥抱大数据的各类行动举措和思维态度, 并分析了大数据这一概念得到多边热议的动机和缘由, 本章下面的结构安排如下: 2.2 节介绍政治家和政府针对大数据时代的来临所做的一系列举措, 或许所有的政策和举措都是因为政府期望通过大数据这个支点能够更好地进行宏观调控、国家治理, 有效减少社会运行成本, 提高经济与社会运行效率, 再或者加快产业结构调整 and 升级, 催生新产业, 带来经济增长新空间; 2.3 节介绍工业界针对大数据时代来临的一系列响应, 包括新型应用模式、商业模式的构建、创新性产品的设计等, 业界如此关心或许是因为工业界的各个环节都认为大数据能够带来更多的商业机会, 也能够成为企业的竞争力源泉, 从而立于商战不败之地; 2.4 节简单介绍学术界对大数据的关注和动作。学术界的关注或许是源于这样的一种价值观——大数据给科学研究、应用研究、工程开发带来的一系列挑战, 作为理性的研究人员, 理当站在迎接这种挑战的一线, 或许也正因为学术界的关注, 所有的困难和挑战就成为一种机遇, 比如“数据科学”这一新型研究范式或学科的构建; 2.5 节对本章进行小结。

2.2 政界大数据

作为大数据的策源地和创新引领者, 美国大数据发展一直走在全球最前列。2012 年 3 月, 美国奥巴马政府发布“大数据研究和发展倡议”, 推进从大量的、复杂的数据集合中获取知识和洞见的能力。该计划涉及美国国家科学基金会 NSF、美国国家卫生研究院 NIH、美国能源部 DOE、美国国防部 DOD、美国国防部高级研究计划局 DARPA 和美国地质勘探局 USGS, 投资总共超过两亿美元, 这笔巨额的科研经费被用于相关工具与技术的开发, 主要包括:

1) 美国国家科学基金会和美国国家卫生研究院主要推进大数据科学和工程的核心方法及技术研究, 项目包括管理、分析、可视化以及从大量的多样化数据集中提取有用信息的核心科学技术。

2) 美国能源部试图通过先进的计算进行科学发现, 提供 2500 万美元基金来建立可扩展的数据管理、分析和可视化研究所。

3) 国防部高级研究局项目主要推进大数据辅助决策, 集中在情报、侦查、网络间谍等方面, 汇集传感器、感知能力和决策支持建立真正的自治系统, 实现操作和决策的自动化。

4) 美国地质勘探局通过给科学家提供深入分析的场所和时间、最高水平的计算能力和理

解大数据集的协作工具,催化在地理系统科学方面的创新思维等。

大数据倡议计划提出之前,事实上在美国已经有许多关于数据驱动的工作陆续展开。从2009年开始,美国联邦政府就开始公开大量资料库,并且把许多数据公布在中央信息交换库——Data.gov网站上,以方便民众进行查阅。联邦政府公布的这些数据给企业进行新的产品、服务开发提供了优质资源,也给民众带来了更好的政府服务,如气象服务、定位服务等。奥巴马在2013年5月签署了第13642号总统行政令,要求在保护隐私安全与机密性的前提下,将数据公开纳入政府的义务范围。

从2010年开始,联邦政府开展一系列主题为“我的大数据”的活动,使美国人可以更安全地获取个人数据来办理私人业务,其中包括:“蓝纽扣”计划(消费者使用“蓝纽扣”从健康服务企业、医药实验室、零售药房供应商与州免疫信息数据库获得他们所需的个人健康信息,以便管理其健康、经济状况,并与信息提供方交换信息)、“创建副本”计划(2014年,美国国税局建立了一个名为“Get transcript”的共享数据库,纳税人在此可以获得个人近三年的纳税记录,使得居民可以方便地下载纳税申报单,更加便捷地进行抵押、贷款等活动)、“绿纽扣”计划(美国政府与电力行业在2012年合作推出“绿纽扣”计划,为家庭与企业提供能源使用信息,目前已为5900万家庭与企业提供服务,并帮助他们节约能源)、“我的学生数据”(美国教育部将助学金免费申请表与联邦助学情况的信息进行共享,这些信息囊括了借贷、补助金、注册与超额偿付等,使学生与资助人能够上网下载所需信息资源)。

美国白宫发布的《2014年大数据白皮书》中提到:“大数据的爆发带给政府更大的权利,为社会创造出极大的资源,如果在这一时期实施正确的发展战略,将给美国以前进的动力,使美国继续保持长期以来形成的国际竞争力。”今天的美国,大数据技术正在催生各个领域的变革力量,从政府到企业,从医疗、教育等公共服务部门到商业、科技领域,整个社会也在不遗余力地主动进行大数据技术的发展与应用。

在美国提出“大数据研究和发展计划”的2012年,我国政府也批复了“十二五国家政务信息化建设工程规划”,专门进行人口、法人、空间、宏观经济和文化等五大资源库的建设工程,总投资额估计有几百亿人民币。2012年3月,我国科技部发布的《“十二五”国家科技计划信息技术领域2013年度备选项目征集指南》把大数据研究列在首位。2014年政府两会工作报告特别提出“设立新兴产业创业创新平台,在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进,引领未来产业发展”。工信部2014年国家物联网重大应用示范工程、发改委2014年4G专项、国家自然科学基金委2014年重点基金、科技部2015年国家科技支撑计划及863项目均专门设立大数据研究、示范应用的项目指南。显然,大数据作为一种国家意志已经上升为国家战略,并且在实施上,社会各界的关注和投入都比曾经的“信息化战略”猛烈得多。

在中央政府高度重视下,地方政府更是积极主动出击,推进大数据服务平台的落地。广东省率先启动大数据战略推动政府转型,北京市正积极探索政府公布大数据供社会开发的可

行性，上海市也启动了大数据研发三年行动计划。在政府部门数据对外开放、由企业系统分析大数据进行投资经营方面，上海市无疑是先行一步。2014年5月15日，上海市推动各级政府部门将数据对外开放，并鼓励社会对其进行加工和运用。根据上海市经信委印发的《2014年度上海市政府数据资源向社会开放工作计划》，目前已确定将190项数据内容作为2014年重点开放领域，涵盖28个市级部门，涉及公共安全、公共服务、交通服务、教育科技、产业发展、金融服务、能源环境、健康卫生、文化娱乐等11个领域。其中市场监管类数据和交通数据资源的开放将成为重点，这些与市民息息相关的信息查询届时将完全开放。这意味着企业运用大数据在上海“掘金”的时代来临，企业投资和上海民生相关的产业如交通运输、餐饮等，可以不再“盲人摸象”。

2010年，欧盟正式发布“欧洲数字化议程”，旨在建立一个统一的“数字市场”，推动欧盟内部高速和极速互联网的互联互通和应用共享，进而促进欧盟经济社会可持续发展，造福欧盟人民。2012年，欧盟委员会在“欧洲数字化议程及其挑战”中制定了大数据战略，强调了公共数据安全及挖掘公共机构数据的价值潜力，满足了民众日益强烈的对个人数据安全保护的诉求，同时发展物联网，确保网络安全及在线交易的数据处理安全。

英国是最早推进大数据规划的欧洲国家。2004年，英国设立了水平扫描中心（HSC）项目，以提升政府处理跨部门和多学科挑战的能力。2011年，水平扫描中心（HSC）启动气候变化的未来国际影响计划，通过对多数据源进行深度分析，研究解决气候变化对食品和水的可获得性以及地区或国际形势的影响等问题。英国政府发起的另外一个项目是2009年建立的 <http://data.gov.uk> 公共网站，来自七个政府部门的1000多个既有数据集对外开放，后来增到8633个数据集。

荷兰、瑞士、英国和其他17个国家与IBM合作开发了一个名为DOME的超级计算机系统项目（该大数据项目的总部位于英国曼彻斯特的Jodrell Bank天文台），该系统每天能处理超过1EB的数据，数据来源于射电望远镜平方公里阵列（SKA）。该系统旨在通过探索百亿亿次的计算、数据传输和存储等新兴技术以及对每日采集的数据流进行读取、存储和分析，解决一系列宇宙科学问题。

2011年，韩国总统国家ICT战略委员会（该委员会是最高层次的政府信息通信技术协同机构，其使命是在建立必要的基础设施过程中发挥领导作用）发布了“大数据倡议”。该倡议旨在建立泛政府大数据网络和分析系统，推进政府与私有部门之间的数据共享融合，建立公共数据诊断系统，培养和培训合格的大数据专业人员，保障个人信息安全以及改善相关法律，发展大数据基础设施和技术，发展大数据管理和分析技术。

在韩国很多政府机构已经提出了相关的行动计划。例如，韩国卫生部建立了社会福利综合管理网络，分析来源于35个机构的385个不同类型的公共数据，综合管理中央政府和地方政府提供的福利和服务。食品、农业、森林与渔业部、公共行政与安全部（MOPAS）计划推出预防手足口病的综合系统，该系统依托于分析动物疾病相关的海外大数据、海关出入境记

录、养殖场的跟踪调查、牲畜迁移和养殖工人活动等相关的大数据,达到预防目的。MOPAS的另一项计划是推出灾害预报系统,该系统基于过去的灾害记录和自动实时的天气和地震预报进行预测。此外,韩国生物信息中心计划开发和运营国家DNA管理系统,该系统集成大量的DNA和病人医疗信息,为个人提供个性化定制的诊断和治疗。

世界上其他发达国家的政府部门已经开始推广大数据应用,如新加坡、日本和澳大利亚均在大数据时代做出了自己相应的响应和行动。通过分析和比较这些先进发达国家的大数据应用,我们能了解当前以及未来需要大数据应用聚焦和服务的地方,并为我们开展大数据应用提供借鉴。

2004年,针对国家安全、传染病和国家层面关心的其他问题,新加坡政府与国家安全协作中心合作发布了风险评估和水平扫描计划(RAHS)。通过对大数据的采集和分析,积极把控制威胁国家安全的相关问题,包括恐怖袭击、传染病传播和金融危机等。风险评估和水平扫描计划实验中心(REC)于2007年开放,它聚焦于风险评估和水平扫描计划相关政策制定的新技术工具,并通过大数据基础设施系统升级来维持和强化这一能力。为通过大数据研究、分析和应用创造价值,新加坡政府还推出了门户网站<http://data.gov.sg>,50多个政府部门的5000多个数据集通过此网站向公众开放。

日本政府已启动多个利用既有大数据的计划。从2005年到2011年,文部科学省与相关的大学和研究机构合作,设立了信息爆炸时代的新IT基础设施项目。从2011年起,政府优先解决地震、核电站灾难和受污染区域的重建和灾民安置以及相关的社会和经济救济等问题。文部科学省与国家科学基金会合作,提高研究和利用大数据的技术,以预防、减轻和管理自然灾害。作为内务省的两个分支机构,信息和通信委员会和ICT战略委员会,把“大数据应用”作为日本面向2020年的关键使命。

澳大利亚政府信息管理办公室(AGIMO)实施政府2.0计划,为公众获取政府数据提供了渠道。政府2.0计划推出了<http://data.gov.au>网站,通过这一网站,让公众便捷、高效地检索和获取政府数据。

对于政府部门来说,大数据将提升电子政务和政府社会治理的效率。大数据的包容性将打破政府各部门间、政府与市民间的边界,使得信息孤岛现象大幅削减,数据共享成为可能,政府各机构协同办公效率和为民办事效率提高,同时大数据将极大地提升政府社会治理能力和公共服务能力,不断拓展个性化服务,进一步增强政府与社会、老百姓直接的双向互动、同步交流。“大数据”的应用已成为当今政府提升执政能力,改善公共服务的重要手段和必由之路。

预测医学是大数据在健康领域的终极运用,这项强大的技术可以同时深入解析一个人的健康状况与遗传信息,使医生更好地预测特定疾病在特定个体上是否可能发生,并预测患者对于特定治疗方式的反应。

在大数据与医疗保健服务相结合方面,美国议会出台了鼓励医疗保健服务供应商使用电

子病历的法案，极大地提高了可供临床医生、研究者与病人使用的数据量，形成一个“学习型”的医疗保健系统，在此系统内临床数据将迅速反馈给患者并有效地指导治疗。

大数据也给教育创新提供了一些新的方案和思路。

在政府的指导下，一些学校通过提供教学效果的实时评估以及精确跟踪在线课堂等学习平台上获取的数据，对学生学习轨迹进行更准确、广泛的研究，深入了解学生在学习活动中的接收效果，根据不同的学习目标，选择合适的学习材料，提高学生个体的学习效果。目前，我国教育部正在研究如何运用这些科技，已开始整合国家教育技术计划下在线教学平台所产生的数据，并计划成立虚拟学习实验室。

大数据带来执法手段的革新，监控（本质是数字化）及存储成本的大大降低给执法部门的数据收集和记录监控提供了极大的便利，同时为执法部门电子证据采集、情报研判、犯罪预测等提供了丰富的数据基础。

洛杉矶与孟菲斯警方所使用的犯罪预测软件“PredPol”的主界面是一张城市地图，它会根据某一地区过往的犯罪活动统计数据，借助各种算法，计算出某地发生犯罪的概率、犯罪类型以及最有可能犯罪的时间段。这样就可以找到需要提高警惕的犯罪“热点”地，在“热点”范围内加强巡逻警力，有效降低了辖区内的犯罪数量。如今这种预测分析技术还能被用于对某一独立个体的犯罪倾向进行分析，对某些特定的人群提高防范。

在讲述奥萨马·本·拉登（Osama bin Laden）丧命经过的《终结》（The Finish）一书中，作者马克·鲍登（Mark Bowden）写道，帕兰提尔（Palantir）公司的软件是“名副其实的杀手级应用”，在美国击毙本·拉登的行动中发挥了情报分析的作用。在过去五年里，帕兰提尔公司已经变成了进行大规模数据挖掘以供美国情报及执法部门使用的关键公司，其软件产品有着流畅的界面，旗下程序员甚至会空降到客户的总部进行程序定制。帕兰提尔公司把混乱无序的大量信息变成直观的可视化地理分布图、柱状图和关联图。只要给该公司所谓的“前沿部署工程师们”几天时间，让他们分析、标记和整合所有零碎的客户数据，帕兰提尔公司就能弄清楚各种各样的问题，如恐怖主义、灾难响应和人口贩卖。

政府使用大数据是为了提升和改善公共服务，这与企业利用其追求利润异曲同工。不同点在于：政府追求的是全民意义上的福祉，而企业追求的是企业法人意义上的获益最大化。政府运用“大数据”的主要目标是维持国内稳定，实现可持续发展，确保公民的基本权利，改善国民福利和促进经济增长。此外，政府的决策制定过程通常需要很长的时间，经过不同群体（包括官员、利益集团和普通民众）反复讨论和磋商，在彼此间达成一致后才能有最终结果。由此看来，大数据在政府部门和私有公司的应用具有很大不同。

政府开展大数据应用，应重视和认识的几个问题包括：

1) 国家优先发展战略：国家需要将大数据产业作为优先发展战略，利用大数据带动其他产业的发展，以提高国家运转效率、透明度、民众福利和公共事务参与度，确保经济增长和

国家安全。

2) 统筹分析机构：对于跨部门的数据，管理和综合数据需要一个自上而下的统筹。政府应建立一个大数据控制中心以综合各部门既有数据的数据库，包括结构化的和非结构化的。此外，政府还需要建立一个先进的分析机构负责开发战略，处理大数据如何通过新技术平台进行管理和分析、如何招募到熟练的从业人员等问题。

3) 实时分析数据：政府需要管理即时更新的大数据，并进行实时分析，同时保护个人信息安全，还需要探索新技术平台（比如云计算、先进分析和安全技术）。相当多的政府数据在性质上是全球化的，而且能够被用于预防和解决全球事务，因此必须开展全球合作。

4) 国际化：各国政府间努力集成和共享地球观测数据。全球地球观测系统是一个全球性的公共基础设施，产生了综合的、接近实时性的环境数据，目的是为全球使用者和决策制定者提供信息以供分析。政府也需要共享与安全威胁、诈骗和非法活动相关的数据。这种大数据需求不仅需要转换技术，还需要国际化的协作去共享和综合数据。

5) 与专业公司合作：例如，亚马逊 AWS (Amazon Web Services) 关联很多公共数据集，包括日本、美国人口调查数据和许多基因组及医疗数据库。

2.3 业界大数据

企业从来是站在科研成果与应用相结合的一线，一方面企业是用户需求的响应者，另一方面也是关键技术、相关理论的研发者和贡献者，这在大数据时代也不例外，比如 Apache Nutch 对大数据的理解是：更新网络搜索索引需要同时进行批量处理或分析大量数据集；而 Google 发布的 MapReduce 和 Google File System 对大数据从数据量级和处理速度量级进行了再定义。事实上，MapReduce 已经成为很多需求场景开展大数据应用的事实标准；Gartner 关于大数据的描述是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产；IBM 对大数据的理解可以归纳为：在掌握信息 (Align) 的基础上获取“洞察 (Anticipate)”，进而采取“行动 (Act)”，不断“学习 (Learn)”，实现“转型 (Transform)”。每个独立的公司根据自身的成长基因都对大数据给予了不同的理解和各自的响应，事实上，最激动人心的是各个公司开发出一系列创新应用。

Google 翻译系统是大数据应用中一个津津乐道的经典案例：

谷歌翻译系统为了训练计算机，会吸收它能找到的所有翻译。它会从各种语言的公司网站上去寻找联合国和欧洲委员会这些国际组织发布的官方文件和报告的译本。尽管其输入源很混乱，但较其他翻译系统而言，谷歌的翻译质量相对而言还是最好的，而且可翻译的内容更多。谷歌的翻译之所以更好并不是因为它拥有一个更好的算法机制，而是谷歌翻译将语言视为能够判别可能性的数据，而不是语言本身。通过有效利用这些散布于互联网的、混杂的数据，将其作为“训练集”，可以正确地推算出英语词汇搭配在一起的可能性。谷歌公司的人工智能专家彼得·诺维格 (Peter Norvig) 和他的同事合作的论文《数据的非理性效果》(The

Unreasonable Effectiveness of Data) 中写道:“大数据基础上的简单算法比小数据基础上的复杂算法更加有效。”

作为网络售书的鼻祖,亚马逊不仅为类似售书的电商提供了可借鉴的解决方案,而且在云计算时代,还通过设备租赁方式提供虚拟化计算资源,成了云计算的一个标杆,并以企业云闻名于世。在大数据时代来临的今天,亚马逊在大数据方面的工作也是可圈可点。

业界戏称(也或许是竞争同行的担忧):亚马逊已经从传统的电商公司进化转型为大数据公司了。亚马逊网站上发生的所有行为都会被亚马逊记录,如搜索、浏览、打分、点评、购买、使用减价券和退货等。亚马逊根据这些数据,不断勾画出每个用户的特征轮廓和需求;同时对用户行为进行整合,亚马逊通过各种交互手段去获取用户的喜好和需求。比较典型的活动就是投票,一旦用户投票了,其观点、倾向或者兴趣爱好就暴露了,这个用户就被亚马逊打上了“标签”;紧接着亚马逊通过对所获行为信息的分析和理解,制定对客户的贴心服务及个性化推荐。这不仅可以提高客户购买的意愿,缩短购买的路径和时间,在恰当的时机捕获客户的最佳购买冲动,降低了传统营销方式对客户的无端骚扰;更重要的是,亚马逊还会对推荐之后的用户行为数据进行统计分析,比如给目标用户发送邮件后,用户是否打开了邮件、是否点击了邮件中的链接浏览促销产品等,这些行为都会被持续跟踪记录下来。所有这些后评估结果为下次类似促销的活动提供了历史依据。

事实上,亚马逊除了针对自身的电商业务充分利用大数据获得更多收益以外,其自主开发的在线分析引擎 Elastic MapReduce 一直很受欢迎,它帮助客户挖掘当前未充分利用的大数据源,然后利用 BI 展示。在 Elastic MapReduce 的基础上,亚马逊还研发了两款全新的大数据服务 DyamODB 和 RedShift,前者是驱动亚马逊消费网站的 NoSQL 产品,后者则是在线数据仓库服务。

如今大数据的价值已得到互联网公司的认可和重视,大规模数据的收集和应用也已经开始为互联网公司的企业决策和营销活动提供强有力的支持。近年来,众多中国互联网企业相继开始了自己的大数据应用,并推出了基于大数据的精准营销服务解决方案。

截至2015年6月30日,腾讯QQ活跃账户为8.434亿、微信活跃账户为6亿、QQ智能终端月活跃账户6.27亿、QQ空间月活跃账户6.592亿、腾讯新闻活跃用户2.8亿……腾讯无疑是目前中国乃至全球最大的互联网综合服务提供商之一。也是中国乃至全球服务用户最多的互联网企业之一。在积累了个人用户多方面的海量数据后,2012年腾讯即提出了“大数据营销”的概念,尝试从这些海量数据中挖掘、分辨出用户的行为模式、兴趣偏好等,打造专属于每个人的智慧门户。腾讯不仅在各大产品线中都设置了数据挖掘团队,还和一些第三方数据挖掘公司、营销公司展开合作洽谈,充分挖掘用户在网上行为、关系、UGC(用户产生的内容)等数据。腾讯的一个商业目标是:通过合理的方法找到对企业有帮助的数据,并且将营销预算合理地分配在为数众多的数据来源平台上,从而提高营销效率。

事实上,在大数据时代的大数据一线从不缺少中国企业的身影,除了腾讯、百度、阿里巴巴等互联网航母公司及许多创新型、成长型大数据公司均在不断推出基于大数据的精准营销服务解决方案,此处不再赘述。

电信运营商为摆脱“管道化”的困境,也纷纷在大数据方面挖掘价值潜力。电信与媒体市场调研公司 Informa Telecoms & Media 在 2013 年的调查结果显示,全球 120 家运营商中约有 48% 的运营商正在实施大数据业务,这是市场倒逼使然:随着移动互联网时代的到来,用户的消费习惯及行为习惯均发生了很大的改变,从任何一家电信运营商的营业财报来看,其语音业务及短信业务逐年下降的同时,数据业务大幅提升,如果不能在数据业务中获取更多的市场份额和盈利份额来补充在语音业务及短信业务上的短板,势必会在恶劣的竞争环境中被淘汰出局。但是电信运营商的固有优势也很明显:

1) 运营商为移动互联网的迅猛发展提供了几乎无法复制的通信平台。作为流量的入口,任何一家电信运营商都拥有海量的用户基础以及这些海量用户每天都在产生的海量数据。

2) 电信运营商拥有多年的数据积累,拥有诸如财务收入、业务发展量等结构化数据,也会涉及图片、文本、音频、视频等非结构化数据。

3) 从数据来源看,电信运营商的数据来自于移动语音、固定电话、固网接入和无线上网等所有业务,也会涉及公众客户、政企客户和家庭客户,同时也会收集到实体渠道、电子渠道、直销渠道等所有类型渠道的接触信息。

如何充分采集、整合、有效利用这些数据,发现数据背后隐藏的信息,从而为运营商的公共基础设施建设、高效运营、创新应用、服务改良提供技术支撑是每个运营商均要面对的需求痛点。

目前国内运营商运用大数据主要有五个方面:

1) 网络管理和优化,包括基础设施建设优化和网络运营管理优化等。

以基站建设布局及优化为例,通过对用户行为轨迹的分析,可以主动获悉在某个地区内各个基站的人群分布以及各个基站的通信质量负载,借此可以在基站建设和布局优化方面进行主动的辅助决策;通过对用户从各个渠道的反馈数据(比如打电话给客服投诉某个地区通信质量不好)能够有靶向地定位不同基站的网络质量,借此实现基站的管理优化,相比较于传统单一通过工委部门的日常巡检监测而言,这种类似“众筹模式”的数据收集手段更容易获得对具体小区基站的质量评估。

2) 市场与精准营销,包括客户画像、关系链研究、精准营销、实时营销和个性化推荐等。

以移动终端机的推荐为例,电信运营商通过对用户消费数据(不仅是具体某个个体,包括整个区域的用户消费行为和习惯)分析,能够对用户何时换手机、会换什么样的手机等进行精准预测,这样在合适的时间(通过对整个地区的消费者的换机行为进行“用户-使用时间”进行建模分析出具有某属性特征的用户使用当前手机的一般使用时间,国内的一般做法

是在可能换手机的前三个月)、通过合适的渠道(根据消费者的消费习惯建模分析出其渠道偏好)、在合适的时机(比如用户在办理某个相关业务、运营商在进行某个营销案等)向合适的人推荐合适的手机(比如根据用户历史使用手机的偏好建模、根据整个地区手机更换的关系模型等)。

3) 客户关系管理, 包括客服中心优化和客户生命周期管理。

以客户忠诚度评估为例, 运营商对客户忠诚度评估的最终目标是通过用户市场倒逼企业提高服务质量, 最终提高用户的满意度。简单的策略是: 一方面通过对用户的消费流水、消费行为、用户投诉反馈等数据对“用户-选用产品”进行立体化建模; 另一方面通过对企业产品舆情和竞品舆情(各种渠道采集, 比如问卷调查、电话回访、互联网数据采集等)分析, 借此评估普适用户对相关产品、服务和渠道的体验度和满意度并进行建模; 通过对上述所有数据的后评估达到对用户的忠诚度评估。

4) 企业运营管理, 包括业务运营监控和经营分析。

以渠道价值度和健康度评估为例, 为了拓展业务, 各个运营商一般设立了包括自营和加盟的营销渠道, 每个营销渠道都可以为用户办理相关业务。所有用户办理业务时, 后台记录的业务流水能够反映具体每位用户的业务办理情况, 也能反映出这笔业务是哪个区域、哪个渠道、哪位经手人以及是在哪位销售经理的影响下开通的。这样通过对业务流水的分析就可以为相关的人打上各种标签(当然还需要其他数据源数据的支撑), 借此可实现具体用户的价值度分析、渠道价值度分析、相关人的绩效分析等, 同时通过具体某个渠道的业务办理的上下文(如用户办理的相关业务)可以对违规的业务结构进行预警, 借此实现对渠道、相关人的健康度进行分析建模, 从而为运营商的运营管理和有效稽查提供数据支撑。

5) 数据商业化指数据对外商业化, 单独盈利。

以精准广告营销为例, 通过对用户(包括用户朋友圈)消费行为、消费能力、行动轨迹的有效分析, 可以从不同的维度对用户打上各类标签, 同时对待推送广告与用户的上述数据进行关联建模, 达到在合适的时间、合适的区域、用户进行相关的行动时向用户推送合适广告的目标, 从而实现运营商主营业务以外的增值获益。比如某个用户在周末会获得某个餐饮或者旅游点的广告, 可能的原因是通过对用户轨迹历史数据的分析, 推断出这位用户在周末的娱乐圈(地域)一般集中在哪些地方, 消费行为一般集中在哪些等。

当然, 运营商大数据应用的场景肯定不限于此, 这里仅给出一般的应用示意。应该说, 从整体来看, 电信运营商大数据发展仍处在探索阶段。这对于大数据产业链上的不同角色或许都是一个机遇。

如前所述, 大数据市场规模巨大, 获益期望巨大。大数据技术已经渗透进我们社会生活中的方方面面, 通过各种各样的方式影响着社会的发展, 而大数据环境下需要响应的需求逐渐趋于个性化、扁平化、垂直化等, 这就要求无论在应用模式还是商业模式上均需有针对性

地响应。事实上,我们已经发现很多创新应用和创新商业模式,各个公司利用自身的个体资源和成长基因在整个渐趋成熟的大数据产业链中占有各自的位置。按一般的理解,在大数据产业链中有三类典型的公司:

1) 基于数据(本身)的公司:专指那些拥有数据但往往不具有数据分析能力的公司,这类公司往往规模尤其巨大、资源获取能力尤其巨大,一般中小型公司没有绝对的竞争力。

2) 基于技术(研发)的公司:专指那些技术供应商或者数据分析公司等,同基于数据的公司类似,这类公司往往规模巨大,技术研究底蕴深,一般中小型公司没有绝对的竞争力。

3) 基于思维(服务)的公司:专指挖掘数据价值的大数据应用公司,这类公司往往自己不具有大量的数据基础,甚至没有专门的技术理论研发能力,但他们能够清晰地理解目标应用的需求,并且能够嫁接相应的技术给予应对。与前两者的公司比较而言,这类公司往往规模无须巨大(当然也有航母级公司针对行业应用做解决方案),也可开展相应的工作。这类公司需要具备两类基本的素质,即对需求的极致敏感(能够发现、发掘甚至引导和创造用户的需求)、对技术选型的极致敏锐(能够快速选择合适的技术响应相应的需求)。

上述的分类比较粗放,因为很多公司是兼具两个甚至全部的特征。以 Google 为例,作为搜索引擎航母,它首先是一个基于数据的公司;同时它也是引导大数据技术研发的领军企业,其发布的 MapReduce 等已经成为业界的事实标准;另外它还出于企业战略开展了若干面向具体应用(或者抽象层次更高一点)的数据分析服务或系统。就这个意义上而言,Google 是基于数据、技术与思维的公司。

事实上,作为正处于战略转型期和产业结构调整优化的中国产业生态,还有可称为第四类的大数据公司,这类公司往往是处于转型期的供给侧生产型公司本身,出于企业生存和长期获益的追求,尝试利用大数据技术改造既有的生产工艺和流程优化,而中国政府也从国家战略角度给予了支持和引导。

2014 年中国政府工作报告中的第六点“以创新支撑和引领经济结构优化升级”指出:促进信息化与工业化深度融合,推动企业加快技术创造……设立新兴产业创业创新平台,在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进,引领未来产业发展。

2015 年年末,“供给侧”及“供给侧改革”这样比较“学术范”的经济学技术名词因为中国政府高层的密集发声,而成为老百姓街头巷尾热议的关键词。2015 年 11 月 10 日,习近平主席在中央财经领导小组会议上首次提出“供给侧结构性改革”,并在次日召开的国务院常务会议再次强调“培育形成新供给新动力”;2015 年 11 月 17 日,李克强总理在“十三五”《规划纲要》编制工作会议上强调,在供给侧和需求侧两端发力促进产业迈向中高端;2015 年 11 月 18 日,习近平主席在 APEC 会议上发表主旨演讲时,再次提及供给侧改革,指出要解决世界经济深层次问题,单纯靠货币刺激政策是不够的,必须要推进经济结构性改革,使供给体系更适应需求结构的变化。

所谓“供给侧结构性改革”,指的是从供给侧入手,针对经济结构性问题而推进的改革,

传统的需求侧管理，强调的是政府的宏观调控作用、熨平短期的经济波动，在实际操作上更多是宏观经济政策的调整；而供给侧改革，强调的是企业和创业者作为市场主体的作用、解决经济中长期的结构性问题，在实际操作上更注重效率的提升、制度的变革和完善。

事实上，不仅中国政府，或许所有的发达国家、发展中国家都有类似的鼓励本土企业从全局的产业结构调整和企业自身发展需求驱动进行生产工艺的改造倡议和引导。比如出于美国制造业回归目的的振兴美国制造业计划、欧盟推进的新工业革命、德国推进的工业4.0，当然也包括中国政府多年来一直推进的两化融合国家战略，都是从国家意志的角度，针对制造业企业进行的一个战略引导，而大数据是达成这些战略意图过程中不可避免的重要内容，后文会有专门章节加以分析。

将自己定位成“大自然搬运工”的农夫山泉，在全国有十多个水源地。农夫山泉把水灌装、配送、上架，一瓶超市售价2元的550ml饮用水，其中0.3元花在了运输上。在没有数据实时支撑时，农夫山泉在物流领域花了很多冤枉钱。比如某个小品相的产品（350ml饮用水），在某个城市的销量预测不到位时，公司以往通常的做法是通过大区间的调运来弥补终端货源的不足。华北往华南运，运到半道的时候，发现华东实际有富余，从华东调运更便宜。但很快发现对华南的预测有偏差，华北短缺更为严重，华东开始往华北运。此时如果太湖突发一次污染事件，很可能华东又出现短缺。这种没头苍蝇的状况让农夫山泉头疼不已。此外，针对日常运营产生的销售、市场费用、物流、生产、财务等数据，传统的做法是“ETL工具导入数据库→数据关联和展现”，由于数据源彼此独立在不同的业务系统中，物流、资金流和信息流有效汇聚到一起并彼此关联形成一份有价值的统计报告需要24小时，当农夫山泉的每月数据积累达到3TB时，这样的速度导致农夫山泉每个月财务结算都要推迟一天。更重要的是，农夫山泉的决策者们只能依靠数据来验证以往的决策是否正确，或者对已出现的问题做出纠正，仍旧无法预测未来。在采购、仓储、配送这条线上，农夫山泉特别希望基于大数据解决三个痛点：①解决生产和销售的不平衡，准确洞悉应该生产多少、配送多少；②让400家办事处、30个配送中心能够纳入体系中来，形成一个动态网状结构（而非简单的树状结构）；③让退货、残次等问题与生产基地实时连接起来。

2011年6月，SAP和农夫山泉开始共同开发基于“饮用水”这个产业形态中运输环境的数据场景应用。目标是作为神经末梢的销售终端的任何一个痛点能快速感知并反馈到企业大脑中，其中的一个核心问题是如何根据不同的变量因素来控制自己的物流成本，SAP团队和农夫山泉团队纳入很多数据：高速公路的收费、道路等级、天气、配送中心辐射半径、季节性变化、不同市场的售价、不同渠道的费用、各地的人力成本甚至突发性的需求（比如某城市召开一次大型运动会）等。另外一个核心问题是，在数据尽可能全的基础上，如何快速地分析数据。

SAP利用其自主研发的创新性数据库平台SAP HANA，结合自主研发的BI工具套件，有效地达成上述目标并在当年9月宣布系统对接成功。农夫山泉在点评选择SAP HANA的缘由

时说：快些，再快些。采用 SAP HANA 后，同等数据量的计算速度从过去的 24 小时缩短到了 0.67 秒，几乎可以做到实时计算结果，这让很多不可能的事情变为了可能。有了强大的数据分析能力做支持后，农夫山泉近年以 30%~40% 的年增长率，在饮用水方面快速超越了原先的三甲：娃哈哈、乐百氏和可口可乐。根据国家统计局公布的数据，饮用水领域的市场份额，农夫山泉、康师傅、娃哈哈、可口可乐的冰露分别为 34.8%、16.1%、14.3%、4.7%，农夫山泉几乎是另外三家之和。

其实，对于 SAP 而言，类似饮用水的应用场景，SAP 并非没有案例。雀巢就是 SAP 在全球范围的长期合作伙伴。但是，欧美发达市场的整个数据采集、梳理、报告流程已经相当成熟，上百年的运营经验让这些企业已经能从从容面对任何突发状况，他们对新数据解决方案的渴求甚至还不如中国本土公司强烈。就这点而言，可以看到中国的企业在大数据这条路上已经走得很远了。事实上，农夫山泉的故事还没有完，还有更多的机会，当然也是挑战。

从 2008 年起，农夫山泉总部要求每个本土业务员每天需要例行公事地在本土的超市拍摄农夫山泉矿泉水是怎么摆放、位置有何变化、高度如何……按照每天每个业务员在每个超市售卖点拍摄 10 张照片、每个业务员每天负责 10 个售卖点计算，这样每个业务员每天线下采集的数据在 10M 左右。农夫山泉全国有 10 000 个业务员，这意味着每天这样的数据是 100G，每月为 3TB。当这些图片如雪片般进入农夫山泉在杭州的机房时，摆在公司面前的一个事实是：守着一座金山，却不知道从哪里挖下第一锹。从农夫山泉的角度来看，自然希望知道：怎样摆放水堆更能促进销售？什么年龄的消费者在水堆前停留更久，他们一次购买的量多大？气温的变化让购买行为发生了哪些改变？竞争对手的新包装对销售产生了怎样的影响？不少问题目前也可以回答，但它们更多的是基于经验，而不是基于数据。另一方面，如果超市、金融公司与农夫山泉有某种渠道来分享信息，如果类似图像、视频和音频资料可以系统分析，如果人的位置有更多的方式可以被监测到，那么摊开在决策者面前的就是一幅基于人消费行为的画卷，而描绘画卷的是一组组复杂的、分布的、异构的、结构化和非结构化并存的“0-1”数据。如何有效利用这些数据进行更精准的管控物流、精准营销、企业决策，或许是下一步摆在企业家面前的大事件。

2.4 学界大数据

由于大数据处理需求的迫切性和重要性，近年来大数据技术已经得到了全球学术界、工业界和各国政府的高度关注和重视。美国和欧洲一些发达国家政府都从国家科技战略层面提出了一系列的大数据技术研发计划，以推动政府机构、重大行业、学术界和工业界对大数据技术的探索研究和应用。我国也有类似的资助计划，比如 2014 年发改委 4G 专项、2015 年国家基金重点基金、863 及 973 均专门设立了大数据专题，省部级科技主管部门也设立了大数据相关的基础研究、应用开发及示范应用项目指南。无论是外国政府的大数据研究计划，还是

国内外大公司的大数据研发,当前最重视的都是大数据分析算法和大数据系统的效率。因此,当工业界把主要精力放在应对大数据的工程技术挑战的时候,科技界开始了着手关注大数据的基础理论研究。

由于大数据技术的特点和重要性,目前国内外已经出现了“数据科学”的概念,即数据处理技术将成为一个与计算科学并列的新的科学领域(详细分析及介绍见9.3节)。著名数据库专家吉姆·格雷(Jim Gray,图灵奖获得者)在2007年的一次演讲中提出,“数据密集型科学发现(Data-Intensive Scientific Discovery)”将成为科学研究的第四范式(前三个科学研究范式分别是实验科学、理论科学和计算科学)。当前对大数据的技术研究大致可以分为:

1) 大数据的复杂性和计算模型。针对数据规模的巨大性及数据内容的复杂性导致计算高度复杂性,研究可以支持高复杂度计算并能保证高效计算性能的计算模型。将复杂的计算任务分而治之及并行化处理是一个简单的响应策略,因此,在大数据的研究中,分布式计算模型一直是一个研究重点。比如Google提出的MapReduce计算模型也已成为业界的标准大行其道。当然还有其他的一些分布式计算模型,比如Spark、Storm等,每一种计算模型都有其各自的优势和劣势以及相适应的应用场景,详细区别将在后文专门做比较分析。分布式计算的基本出发点是任务的分而治之及并行化。还有一种高性能计算的思路是充分发挥计算机本身的性能,比如GPU计算就是充分利用显卡的闲余计算能力对大量数据的重复性计算,具有较大的优势。

2) 大数据的感知与表示。大数据的来源复杂,有的是来源于既有的自营平台,有的来自于数据采集终端,有的来自于互联网等,如何高效采集和整合不同数据源的数据并以合适的形式表示,从而为后续的数据理解和建模提供数据支撑就是本研究方向的重中之重。涉及的研究内容包括ETL、数据交换、数据集成和融合、数据清洗与过滤等。

ETL(Extract-Transform-Load)用来描述将数据从来源端经过抽取(Extract,指的是从数据源抽取所需的数据)、转换(Transform,指的是对数据进行清洗)、加载(Load,指的是按照预先定义好的结构存入目标数据库)至目的端的过程。在大数据场景中,由于从数据源来的数据或许要面向不同的垂直应用,这意味着在数据采集的时候要尽可能保持数据的原样,即传统的ETL在大数据环境下应该改为ELT,即从来源端抽取(Extract)出相关的数据后直接加载(Load)到目标端数据库中,然后根据需要进行清洗或过滤(Transform)。这是大数据落地应用中的一个朴素思维。

从互联网中采集数据的方式有两种:一是通过“物-物交换”或者“钱-物交换”进行的交易;另外一种方式就是通过编制爬虫工具从互联网中把数据爬取下来。编制一个爬虫并不困难,难度在于,如果需要快速、海量采集互联网的数据,如何保证数据爬取效率是一个重要的问题。利用多个爬虫形成一个爬虫联盟进行协同爬取是一种思路,但是这牵涉到多个实体的协同工作问题,这既涉及分布式协同计算问题,又涉及各个实体之间的合作博弈问题,而这些研究方向和内容对于大数据的研究都是大有裨益的。

3) 大数据的内容建模与语义理解。针对数据规模的巨大性导致传统分析手段失效的挑战,研究面向大数据内容建模与理解的数据建模方法。在这个研究领域至少有几个研究思路:①基于特定的分布式计算模型,将传统的分析手法进行改良,使其适合此分布式计算模型,从而达到利用传统分析手法进行内容建模与理解的目的。②研究一系列新算法和新理论,使得针对大数据的内容理解与语义分析的计算复杂度大为降低,从而实现建模与理解目标。③通过一些面向具体应用领域的问题转换,将面向大数据的内容建模与语义理解降级为尺度更低的计算,从而可以通过普通的算法达到既定目标。

4) 大数据的存储与架构体系。针对大数据在数据规模上的巨大性及数据内容的复杂性特点,研究应对上述挑战的数据组织方法及数据存取方法。事实上,关于数据组织与存取的研究往往需要耦合数据存储硬件层次的进展。纵观数据组织的研究进展史可以发现,从早期的文件模式发展为网状数据库(适用于磁带存储模式),再后来发展到关系数据库(适用于磁盘/鼓),随着大数据时代的来临,根据大数据的内容特点,NoSQL作为一个适应性比较强的存储方案得到业界的普遍认同,比如Redis、MongoDB等,关于NoSQL的详细内容,参见6.2.2节。

越来越多的数据中心出于性能和能耗的考虑开始大量地采用固态硬盘产品来取代传统的机械硬盘;SSD固态硬盘优于传统HDD机械硬盘最显著的特性,就在于它拥有极速的读写速度以及超低功耗等优势,同时故障率低也是它的另一大优势,因此众多企业用户,尤其是互联网用户开始使用SSD固态硬盘,这也为固态硬盘市场带来了全新的机遇。

5) 其他相关基础研究支撑。完整的大数据落地应用除了需要上述研究成果的积淀以外,还需要其他一些相关技术的支撑,比如安全保障技术(数据安全、信息安全、网络安全等)、虚拟化方法、计算架构、负载均衡等。

云计算(Cloud Computing)是基于互联网的相关服务的增加、使用和交付模式(本质是一种软件或系统的部署实施模式),这种模式提供可用的、便捷的、按需的网络访问,进入可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件、服务)。这些资源能够被快速提供,只需投入很少的管理工作,或服务供应商进行很少的交互。云计算的核心技术是数据虚拟化和计算虚拟化。大数据和云计算有着天然的耦合性,甚至有人断言“没有大数据的云计算就如同没有住户的房地产”,因此在大数据的研究队伍中,有一部分的研究力量来源于云计算的研究团队。而关于云计算的研究一般又有两个分支来源:一方面是做服务计算的研究者迎合云计算的场景进行的若干改良研究。另一方面是以Hadoop平台为基础做更为稳定的、更可靠的Hadoop改良研究。就这个意义上而言,云计算、服务计算、分布式计算、Hadoop等研究方向或领域都可纳入大数据的研究视角范围内。

其他方面,大数据落地应用的核心是对数据进行加工、分析和呈现,由于数据或许牵涉到个人的隐私或者商业机密,因此数据以及数据的使用必须在安全的环境下进行,因此围绕安全(包括数据安全、信息安全、网络安全)的研究是大数据研究中不可或缺的一部分。其

他方面的相关技术此处不一一赘述。

通过上述分析可以看出,与大数据有关的研究几乎涉及了计算机科学与技术研究中的大部分研究方向,或许也是因为这个原因,所有的研究者都能够在大数据这个概念驱动的研究场景中定位好自己的位置,各个方向的研究共同推进大数据研究的深入和拓展。2008年,《Nature》就专门设立了一个专刊,专门讨论大数据对各个学科的影响、挑战和机遇,下文简单介绍几个大数据与相关基础学科的交叉研究方面的事实:

1) 大数据与脑科学的交叉研究。2012年11月16日,加州大学圣迭戈分校德米特里·戈里尤可夫(Dmitri Krioukov)在《Scientific Report》发表论文“Network Cosmology”,提出互联网与脑神经网络的发展与构造具有高度的相似性。研究组利用计算机模拟并结合多种其他计算,证明在复杂网络的动态发展和控制中,描述大尺度时空结构的因果关系网络的曲线图,是一个具有显著聚类特征的幂函数曲线,和许多复杂网络如互联网、社交网、脑神经网络等有高度的相似性。德米特里·戈里尤可夫的研究对于互联网虚拟大脑的设想给予了有力的数据支持。

2) 大数据与金融学的交叉研究。大数据在金融行业应用范围较广,典型的案例有花旗银行利用IBM沃森电脑为财富管理客户推荐产品;美国银行利用客户点击数据集为客户提供特色服务,如有竞争的信用额度;招商银行利用客户刷卡、存取款、电子银行转账、微信评论等行为数据进行分析,每周给客户发送有针对性的广告信息,里面有顾客可能感兴趣的产品和优惠信息。可见,大数据在金融行业的应用可以总结为以下五个方面(不限于):

①精准营销:依据客户消费习惯、地理位置、消费时间进行推荐。

②风险管控:依据客户消费和现金流提供信用评级或融资支持,利用客户社交行为记录进行信用卡反欺诈研判。

③决策支持:利用决策支撑技术进行抵押贷款管理,利用数据分析报告实施产业信贷风险控制。

④效率提升:利用金融行业全局数据了解业务运营薄弱点,利用大数据技术加快内部数据处理速度。

⑤产品设计:利用大数据计算技术为财富客户推荐产品,利用客户行为数据设计满足客户需求的金融产品。

3) 大数据与教育学的交叉研究。随着技术的发展,信息技术已在教育领域有了越来越广泛的应用。考试、课堂、师生互动、校园设备使用、家校关系……只要技术达到的地方,各个环节都被数据包裹着。在课堂上,数据不仅可以帮助改善教育教学,在重大教育决策制定和教育改革方面,大数据更有用武之地。美国利用数据来诊断处在辍学危险期的学生、探索教育开支与学生学习成绩提升的关系、探索学生缺课与成绩的关系。举一个比较有趣的例子,教师的高考成绩和所教学生的成绩有关吗?究竟如何,不妨借助数据来看。比如美国某州公立中小学的数据分析显示,在语文成绩上,教师高考分数和学生成绩呈现显著的正相关。也就是说,教师的高考成绩与他们现在所教语文课上的学生学习成绩有很明显的关系,教师的

高考成绩越好,学生的语文成绩也越好。这个关系让我们进一步探讨其背后真正的原因。其实,教师高考成绩高低某种程度上是教师的某个特点在起作用,而正是这个特点对能否教好学生起着至关重要的作用,教师的高考分数可以作为挑选教师的一个指标。如果有了充分的数据,便可以发掘更多的教师特征和学生成绩之间的关系,从而为挑选教师提供更好的参考。大数据还可以帮助家长和教师甄别出孩子的学习差距和有效的学习方法。比如,美国的麦格劳-希尔教育出版集团就开发出了一种预测评估工具,帮助学生评估他们已有的知识和达标测验所需程度的差距,进而指出学生有待提高的地方。评估工具可以让教师跟踪学生的学习情况,从而找到学生的学习特点和方法。有些学生适合按部就班,有些则更适合图式信息和整合信息的非线性学习。这些都可以通过大数据收集和分析很快识别出来,从而为教育教学提供坚实的依据。在国内尤其是北京、上海、广州等城市,大数据在教育领域已有了非常多的应用,譬如像慕课、在线课程、翻转课堂等,其中就应用了大量的大数据工具。毫无疑问,在不远的将来,无论是教育管理部门,还是校长、教师以及学生和家,都可以得到针对不同应用的个性化分析报告。通过大数据的分析来优化教育机制,也可以做出更科学的决策,这将带来潜在的教育革命。不久的将来,个性化学习终端将会更多地融入学习资源云平台,根据每个学生的不同兴趣爱好和特长,推送相关领域的前沿技术、资讯、资源乃至未来职业的发展方向等,并贯穿每个人终身学习的全过程。

4) 大数据与气象学的交叉研究。2012年7月21日北京遭遇特大暴雨,在一天之内,平均降雨量达164毫米,这是北京市61年以来最大规模的暴雨。此次暴雨因来势凶猛而给广大市民的生活带来了巨大影响。其实,遇到这种情况最主要的还是需要气象部门及时、准确地做出预警,并协同其他运营商,将这种预警信息第一时间下发给北京市民(包括在京旅行的人士)。也正是如此,这场暴雨不仅暴露出了管理工作上的漏洞,也引起了业内人士一场关于“大数据”的探讨。气象对社会的影响涉及方方面面。传统上依赖气象的主要是农业、林业和水运等行业部门,而如今,气象俨然是21世纪社会发展的资源,并支持定制化服务满足各行各业用户需要。借助于大数据技术,天气预报的准确性和实效性将会大大提高,预报的及时性将会大大提升,同时对于重大自然灾害,如龙卷风,通过大数据计算平台,人们将会更加精确地了解其运动轨迹和危害的等级,有利于帮助人们提高应对自然灾害的能力。天气预报准确度的提升和预测周期的延长将会有利于农业生产的安排。尤其是进入秋冬季以来,我国多个城市爆发雾霾天气,空气污染严重。随着PM_{2.5}对人体健康的危害日益被公众熟知,人们对于“雾霾假”的呼声也越来越高。有人调侃,重度污染天,走在上班路上就是一台“人肉吸尘器”。由此看来,依靠大数据分析北京市或其他城市空气污染的形成及对策,任重道远。一是数据的来源。高耗能企业的生产规模、排放量这些数据是否层层上报,准确统计?掌握此数据的部门是否能向社会公开?北京500万辆汽车所加汽油到底有哪些成分,产生的尾气对空气污染指数的“贡献率”到底多大?二是要冲破数据挖掘分析应用的技术壁垒,当然前提就是数据公开。在美国,NOAA(国家海洋暨大气总署)其实早就在使用大数据业务,每天通过卫星、船只、飞机、浮标、传感器等收集超过35亿份观察数据。收集完毕后,

NOAA会汇总大气数据、海洋数据以及地质数据,进行直接测定,绘制出复杂的高保真预测模型,将其提供给NWS(国家气象局)做出气象预报的参考数据。目前,NOAA每年新增管理的数据量就高达30PB(1PB=1024TB)。由NWS生成的最终分析结果,就呈现在日常的天气预报和预警报道上。

2.5 本章小结

或许整个人类历史上都很少有类似“大数据”这样的概念或者现象同时得到“政产学研商用”各界的普遍关注和重视,并且其影响还处于持续发酵中。

作为一个概念,大数据是赚足了媒体的眼球,几乎每天都有关于大数据的大量新闻,如政策导向解读、创新产品发布、行业趋势分析等。大数据在受到多边关注、重视和投入的同时,也引发了诸多难题。大数据时代来了,你准备好了吗?你有多少数据?这些数据有什么用?这些数据如何才能有用?作为一门技术和学科,大数据也引发了研究者(包括科研院所和企业研究院等)的极大关注,甚至愿意创造出科学研究的第四范式来应对其中的挑战。作为一种思维,大数据一直吸引着聪明的人群去思考:应用在哪里?需求怎样?如何响应需求?而作为一种战略,大数据俨然已经成为各国政府博弈的战场,几乎每个国家政府都在为了自主的国家意志出台各项政策、导向和指南。

本章试图介绍不同角色的实体针对大数据的态度和举措,可以看出“政产学研商用”各界对大数据的关注、重视和热议都有发自其自身角色特点的动机和缘由。

大数据作为“未来的石油”,极大地刺激了各国政府的神经,出于国家战略资源的守护和国家博弈的较量,任何一个政府都无法回避、也必须积极响应大数据时代的到来。因为大数据是建立透明政府、宏观调控、国家治理、社会管理、减少社会运行成本、提高经济与社会运行效率、加快产业结构调整和升级、催生新产业、带来经济增长新空间的信息基础,更重要的是,针对这么重要的战略资源,政府更需要在隐私保护、信息安全、数据主权博弈等方面提出切实的立法举措,并利用国家机器进行务实的监管和保障。

另一方面,大数据除了是一个几乎可以无限炒作的概念之外,它确实能够直接提高企业的竞争力。作为以利益最大化为目标的各个企业,正在以不同的方式拥抱大数据时代的来临。哪怕困难重重,也会奋不顾身,于是渐乎成熟的大数据产业链形成了。产业界对大数据的追捧,其根本的原因在于:已经发生的许多事实表明大数据能够带来巨大的商业机遇的同时,也是企业核心竞争力的源泉。

学术界对大数据的关注和热情,其原因在于:①大数据引发的挑战势必引发思维模式的转变,而这有可能诱发新的理论技术,这是学界研究者的核心兴趣所在。②大数据本身蕴含丰富的待突破的理论技术问题,这会刺激学界的研究者不断前行。③大数据不是一朝一夕或是单打独斗就可以完成预研的,至少需要大量资源的交互,而不同资源的交互势必会促进集体智慧的涌现,这也是学界的研究者所欢迎的。归根到底,大数据能够激励学术界的研究者

的持续跟进,其本质原因或许来源于研究者(甚至是整个人类实践过程中的每一个理性人)一个普适的价值观:针对社会实践中遇到的所有问题,理应从科学上证明和分析其可否计算,从技术上研究和发明如何计算。而在这个过程中无论面临怎样的挑战,作为一个理性的、有责任感的研究者,理应有所担当。

当然,大数据时代不长的历史中也有很多不和谐的声音,比如隐私问题、信息安全问题甚至大数据自身就有的若干原罪问题(见3.4.3节)等。不过,任何事物的发展总是在曲折中前进、螺旋式上升的,大数据应该也不例外……

诗人白居易在《钱塘湖春行》中云:“乱花渐欲迷人眼,浅草才能没马蹄。”从哲学的眼光来看,“浅草”“乱花”都是刚刚出现不久的新事物,还没得到推进发展,但是正因为此才“迷人眼”“没马蹄”,因此不妨全方位地、乐观地、正能量地来看待“乱花”“浅草”等现象。

对待大数据现象,何尝不是如此呢?

本章参考文献

- [1] Agneeswaran V S. Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives [M]. New Jersey: FT Press, 2014.
- [2] Alam S, Chowdhury M M R, Noll J. Senaas: An Event-Driven Sensor Virtualization Approach for Internet of Things Cloud [C]. Networked Embedded Systems for Enterprise Applications (NESEA), 2010 IEEE International Conference on, 2010: 1-6.
- [3] Anderson J P. Computer Security Technology Planning Study [R]. Air Force Systems Command, 1972.
- [4] Cybenko G. Dynamic Load Balancing for Distributed Memory Multiprocessors [J]. Journal of Parallel and Distributed Computing, 1989, 7(2): 279-301.
- [5] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.
- [6] Krioukov D, Kitsak M, Sinkovits R S, et al. Network Cosmology [J]. Scientific Reports, 2012, 2(20): 10272-10284.
- [7] Tiwari S. Professional NoSQL [M]. Manhattan: John Wiley & Sons, 2011.

大数据产业

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的刘勇、陈厚兵、尹康、杨骏元、肖雨奇、王咏乾、冯翰洋、韩建军、王楠等几位同学的协助，在此表示深深的谢意。

3.1 引言

盘古之君，龙首蛇身，嘘为风雨，吹为雷电，开目为昼，闭目为夜。死后骨节为山林，体为江海，血为准渚，毛发为草木。（引自《五运历年纪》）

俗说天地开辟，未有人民，女娲抟黄土做人。剧务，力不暇供，乃引绳于泥中，举以为人。故富贵者，黄土人；贫贱者，引绳人也。（引自《风俗通义》）

女娲祷祠神，祈而为女媒。因置昏姻。（引自《风俗通义》）

盘古开天辟地，于是就有了天和地；女娲抟土造人，于是就有了人类；女娲立婚姻制度，于是就有了人类的繁衍……正如马克思所说：“任何神话都是用想象或借助想象以征服自然力、支配自然力，把自然力加以形象化。”于是在中国，儒家文学、道家文学和民间信仰结合成一体，形成了不同版本的创始神，如天、天吴、毕方、据比、竖亥、烛龙、女娲、盘古、三清和玉皇大帝等；在国外也有上帝造人说（《圣经·创世纪》）、创造力之神哈奴姆和女神哈托造人说（埃及神话）、取火神普罗米修斯造万物说（希腊神话）、创世者庞德·杰尔造人说（澳大利亚）等。

总之，这个世界有了人类，并因为人类的存在，世界也不断地丰富多彩起来，当然在这个过程中，人类本身也在不断地发展和进步。

在人类的整个文明进程中（包括之前的蛮荒时代），人类有别于其他物种的一个“法门”是主动地在生产实践中总结出用于指导未来生产实践的知识和经验，并将这种知识和经验以教学传承的方式延续下来，并不断改进，而这种不断采集和整合的知识和经验无疑成为后来人类拥抱和改造自然界的“法宝”，从而最终使得人类成为世界万物的“主宰”。

在人类从蛮荒走向文明的过程中,语言的诞生、文字的创造、造纸术及印刷术的发明无一例外地彰显了劳动人民的智慧,也进一步把知识的传承向着更为便利、更为普惠的方向发展,人类的文明进程也因此以更快的速度向前推进。

标志着第一次工业革命开始的“蒸汽机”的发明和推广将人类的劳动生产力大大提高,同时社会分工也进一步明细,产生了工匠和工程师的分工,发明家成为一个独立的工种;而电力的发明直接将人类推进到第二次工业革命时代,在这个时代里,自然科学同工业生产紧密结合起来,生产力进一步提高,几个主要的资本主义国家因为较早地进行第二次工业革命改造而较早地进入现代资本主义阵营;二战后,以核能技术、空间技术及信息技术为标志的第三次工业革命(更多的场合,人们更愿意用第三次技术革命来表示)从开始至今,信息技术作为一个普惠全球的重要技术,强有力地推动了人类生产效率、生活质量等以超出历史上任何一个时间段的发展速度快速发展。

现在,我们进入了一个被标签化为“物理信息系统融合”的工业4.0时代,有关工业4.0及工业4.0带来的发展机遇,第13章“大数据机遇”会专门介绍,此处不再赘述。本章关注的是工业4.0所内生的大数据问题,更确切地说,信息化水平的突飞猛进必然引发的大数据问题,当前的产业是否已经有充足的准备去响应和拥抱……

所谓产业,指的是经济社会的物质生产部门,每个部门都专门生产和制造某种独立的产品。计算机产业包括计算机制造业和计算机服务业,后者又称为信息处理产业或信息服务业。伴随着大数据时代的到来,大数据涉及的数据采集、数据分析、数据应用等环节都是大数据产业的一部分。就目前的产业布局来看,大数据解决方案提供商(提供技术支持、运维服务等)、大数据处理服务提供商(提供在线或离线的数据服务及数据分析服务)、数据资源提供商是三种主流的业务模式,这也雷同于《大数据时代》一书中对大数据产业链的公司角色进行的划分:数据型公司(数据是公司的核心竞争力)、技术型公司(技术是公司的核心竞争力)和思维型公司(拓展大数据应用领域,对技术和应用的敏感度是公司的核心竞争力)。

针对大数据这一新生概念(当然,现在看来其实也不新了),不乏充满智慧的创业者,他们及时地把握住了时代的机遇,很快介入大数据的市场,如雨后春笋般开创了一个又一个具有大数据概念标签的创业公司,并都在为成为这个领域(细分市场)的领袖而努力。而传统的业界航母,如微软、IBM、Oracle等也敏锐地发觉了大数据的发展契机(或许也是市场倒逼使然),在既有产品的基础上不遗余力地进行面向大数据产业需求的改良和创新,以便在大数据这个市场环境中继续保持航母领袖的地位……

以微软和谷歌的产品应用模式和商业模式来看:

1) 在PC时代,微软是当仁不让的霸主,操作系统是计算机硬件和应用软件的虚拟平台,微软在操作系统上的绝对垄断性使得其可以左右硬件厂商和软件厂商,比如:硬件厂商的每次硬件升级期望微软发布新型操作系统以彰显硬件的性能,而每当PC销售低迷的时候,硬件厂商也期望微软发布新版操作系统以刺激消费者购买新机器的欲望;而基于Windows的软件

开发则更加被动，几乎 Windows 版本的每次升级，都意味着软件开发的持续迭代。回顾和微软竞争过的一些非互联网企业，几乎都被微软霸权压迫得苟延残喘。微软的商业模式是：公司开发软件，然后用户购买软件的使用权，微软从消费者中获得高额收益后持续进行新产品的研发，然后继续卖给消费者，以此循环，微软也乐此不疲。但是大数据时代的突然到来，操作系统不再是产业链的核心，而边缘化为一个吸引用户使用、采集用户行为的渠道，微软是眼睁睁地看着其他的公司利用微软的操作系统将用户数据录入竞争对手的服务器而无法有更多的作为，在“数据即资产”的大数据时代，以操作系统为核心的微软自然屡屡被动。更可怕的是：移动终端的逐步普及，PC 的市场份额也在不断地被移动终端弱化，在移动终端领域异军突起的 Android、iOS 也在逐步蚕食微软（在移动终端的）操作系统的份额……

2) 与微软依赖操作系统、占领桌面市场的策略不同，谷歌基于的是网页数据和用户搜索数据。谷歌从其诞生起就具有大数据的基因，谷歌的商业模式是：公司免费为消费者开发软件、提供服务，然后通过对消费者数据的收集，将广告商的广告推送给消费者，谷歌通过从广告商处获得收益，以此获益为基础，继续为消费者提供更多、更好的免费服务，继续从广告商处获得收益，以此循环。这种“羊毛出在猪身上狗买单”的商业模式，谷歌、广告商、消费者均乐此不疲，因为彼此均获益。相比较于微软，谷歌的这种基于数据的商业模式无疑具有更大的优势，这也或许是其异军突起并迅猛发展的重要原因。

事实上，利用上述的对比方法，我们还可以继续去比较分析其他的一些航母巨头，比如具有庞大的商品数据及用户购买数据的亚马逊帝国、拥有数十亿用户社交数据的 Facebook 以及拥有通信和移动生活数据的苹果等以数据资产为中心的新型科技巨头等。当然也可以分析传统 IT 航母，比如立足大型机、服务企业的 IBM，立足数据库、介入云计算的 Oracle 等。此处不一一赘述。不过可以断言的是：谁拥有更多的数据，谁就会更容易笑傲江湖。当然，不同公司的不同数据资产，也会决定各个公司未来的走势。

2015 年 7 月 29 日，微软正式发布 Windows 10 操作系统，并且声称“Windows 10 即服务”，即：Windows 10 作为一项服务将永远更新和升级。这种从传统的“操作系统产品为导向”演变为“操作系统服务为导向”事实是市场倒逼使然。一个有趣的现象是在 Windows 10 里有一个新增的功能“Cortana 数字助手”，这项功能的确比较方便，但同时你的语音语调、家庭住址、搜索内容、个人爱好、工作安排、约会时间地点、“朋友圈”（联系人信息）等内容都会记录在案，并且上传到微软服务器。可以说你的一切尽在 Cortana 掌握之中，排除掉隐私问题及数据问题不论（这个问题肯定存在），我们会发现微软通过操作系统这个渠道，有意采集和整合各类数据的动机已经很明显了。

总之，围绕大数据涉及的相关业务，由传统 IT 巨头及新型创业公司共同打造的整个大数据产业链逐渐明晰和成熟。

本章尝试梳理大数据这一概念持续得到多边认同的当下，因为不同的价值期望使然而引发不同利益主体迎接和响应大数据带来的机遇和挑战的各类决策和行动，并给出大数据潜在

应用场景的理性评判依据,本章下面的结构安排如下:3.2节简述整个大数据生态环境,包括政策环境、应用环境和技术环境等;3.3节以目前投行普遍认可的大数据产业地图描绘整个大数据产业链的各个环节的角色定位和技术布局;3.4节简单介绍一些在大数据环境下,以“百家争鸣”形式构建的若干大数据平台及其应用场景;3.5节对本章进行小结。

3.2 大数据产业环境

3.2.1 政策环境

大数据概念得到普遍的追捧和热议,国家和政府层面的战略规划和引导是其中的一个关键因素。奥巴马政府在2012年发布的“大数据研究和发展倡议”直接触动了各国媒体政要的神经,近乎引发了类似战备竞赛一样的、国家层面的大数据博弈。

2012年奥巴马政府发布的“大数据研究和发展倡议”从国家战略角度提出推进从大量的、复杂的数据集合中获取知识和洞见的能力。作为针对此发展倡议的响应,国家层面的至少6个部门拟投资总共超过两亿美元,推动和改善与大数据相关的收集、组织和分析工具及技术;而联邦政府各部门大数据计划也在陆续展开……

中国政府在2014年两会政府工作报告中,首次将大数据纳入国家战略,并在次年的两会政府工作报告中发布“互联网+”国家战略的同时,再次强调对大数据的重视。

出于对国家大数据战略的响应和支持,工信部、国家发改委、国家自然科学基金委员会、科技部陆续发布指南性文件,从政府引导基金及政策导向上加强对大数据项目技术的预研和产业化,以基础理论预研为目标的国家自然科学基金委信息科学部也在2014年发布了围绕大数据进行理论预研的重点基金项目指南“大数据技术和应用中的挑战性科学问题研究”(具体包括:面向大数据的知识表达、推理及在线学习理论与方法、基于认知计算的大数据分析方法、面向大数据的粒计算理论与方法、大数据环境下复杂多媒体内容分析、推送与展示、大数据管理系统评测基准的理论与方法、多层多域网络化大数据的高效传输理论与方法、大数据高效能存储与管理方法、大数据高时效计算体系结构与关键技术、大数据结构与关系的发现与简约计算方法、基于大数据的复杂系统行为预测与控制),在其2015年项目指南当中也有许多大数据相关的项目,如基于大数据和领域知识的复杂化过程能效评价与系统优化、基于大数据和云计算的知识自动化决策系统设计与应用验证、面向大数据内存计算的新型计算机体系结构等。而这些国家部门对口的各个省及直辖市相应垂直部门也陆续发布大数据项目指南……

2015年6月17日国务院常务会议再次强调大数据运用的重要性,提出在重点领域引入大数据监管并通过运用大数据等现代信息技术促进政府职能转变,用政务“云”提升政府服务和监管效率、造福广大群众。该次会议明确提出三点意见:①加快政务信息化工程建设,推

动政府信息公开共享；②推进市场主体信息公示，建设信用信息共享交换平台，大力发展信用服务业；③在环保、食品药品安全等重点领域引入大数据监管，主动查究违法违规行为。2015年7月1日，国务院办公厅印发《关于运用大数据加强对市场主体服务和监管的若干意见》则明确地提出将大数据技术应用于市场监管和政府职能转变中。

《关于运用大数据加强对市场主体服务和监管的若干意见》立足市场监管与政府职能转变这一出发点，提出四项主要发展目标：①提高大数据运用能力，增强政府服务和监管的有效性。高效采集、有效整合、充分运用政府数据和社会数据，健全政府运用大数据的工作机制，将运用大数据作为提高政府治理能力的重要手段，不断提高政府服务和监管的针对性、有效性。②推动简政放权和政府职能转变，促进市场主体依法诚信经营。运用大数据提高政府公共服务能力，加强对市场主体的事中事后监管，为推进简政放权和政府职能转变提供基础支撑。以国家统一的信用信息共享交换平台为基础，运用大数据推动社会信用体系建设，建立跨地区、多部门的信用联动奖惩机制，构建公平诚信的市场环境。③提高政府服务水平和监管效率，降低服务和监管成本。充分运用大数据的理念、技术和资源，完善对市场主体的全方位服务，加强对市场主体的全生命周期监管。根据服务和监管需要，有序推进政府购买服务，不断降低政府运行成本。④政府监管和社会监督有机结合，构建全方位的市场监管体系。通过政府信息公开和数据开放、社会信息资源开放共享，提高市场主体生产经营活动的透明度。有效调动社会力量监督市场主体的积极性，形成全社会广泛参与的市场监管格局。

2015年8月31日，中国政府发布的《国务院关于印发促进大数据发展行动纲要的通知》则进一步从国家意志层面，为中国的大数据战略规划部署持续的发展路径。

《国务院关于印发促进大数据发展行动纲要的通知》立足我国国情和现实需要，从国家意志层面制定了未来5~10年的大数据发展规划，包括**五大目标**（打造精准治理、多方协作的社会治理新模式；建立运行平稳、安全高效的经济运行新机制；构建以人为本、惠及全民的民生服务新体系；开启大众创业、万众创新的创新驱动新格局；培育高端智能、新兴繁荣的产业发展新生态）、**三大任务**（加快政府数据开放共享，推动资源整合，提升治理能力；推动产业创新发展，培育新兴业态，助力经济转型；强化安全保障，提高管理水平，促进健康发展）及**十大专栏**（政府数据资源共享开放工程；国家大数据资源统筹发展工程；政府治理大数据工程；公共服务大数据工程；工业和新兴产业大数据工程；现代农业大数据工程；万众创新大数据工程；大数据关键技术及产品研发与产业化工程；大数据产业支撑能力提升工程；网络和大数据安全保障工程），同时该通知还明确了**七大政策机制**（完善组织实施机制；加快法规制度建设；健全市场发展机制；建立标准规范体系；加大财政金融支持；加强专业人才培养；促进国际交流合作）以保障通知精神的有效落实。

随着互联网基础设施的完善和相关分析技术的成熟，在企业和政府部门需求及政策支持的共同推动下，大数据产业正步入快速发展期，从事大数据采集整合、数据存储运算、数据分析挖掘、数据应用和消费服务等细分领域，将获得市场扩张机遇。而大数据应用的广泛拓

展也必将进一步推动相关产业（比如存储设备厂商、硬件厂商、方案提供商）的联动发展。

3.2.2 应用环境

大数据这一概念得到媒体热炒、多边热议的现状反映出了人们对“大数据”的认可和期待，而这种认可和期待是否会落实直接取决于是否能够将大数据务实地落地并真正创造价值。影响大数据是否能够务实落地的因素包括（不限于）：

- 1) 是否有丰富的大数据应用场景？
- 2) 大数据是否能够响应其中的“痛点”需求？
- 3) 是否有坚实的数据基础？
- 4) 是否有良好的 IT 建设基础？

基于这四个评估指标，我们可以发现，目前大数据所面临的应用环境比任何一个时候都要好，这意味着现在恰是大数据落地的最好时机，以国内现状为例来分析其中的若干利好因素：

1) 2015 年，我国两会政府工作报告发布了凸显中国自信和国家意志的“互联网+”战略，为大数据的发展和推进提供了极好的发展机遇。“互联网+X”或“X+互联网”在表征各行各业拥抱互联网的同时，也在暗示这种“拥抱”的不同阶段或方式（技术融合、产品融合、业务融合、产业衍生）所催生的应用模式、商业模式的演化和创新，在不断延展愈来愈多的应用需求。而政府从国家战略层面倡议的“大众创业、万众创新”事实也在暗示“大众（可以）创业、万众（应该）创新”，这意味着会有更多有智慧的人（自然人或者法人）发掘出更多的应用需求和应用场景，而所有这些需求的内在基础是沉淀在互联网上的大数据。或者说：大数据是现状，也是基础，更是响应需求的手段和切入点。因此大数据应用场景无限广阔（详见第 13 章《大数据机遇》，此处不再赘述）。

2) 2015 年 5 月 8 日，我国政府发布的国家战略《中国制造 2025》，事实上是我国政府一直倡导的两化融合在“互联网+”新时期的推进和演化，其本质是通过工业互联网的有效推进达到产业结构调整的目的，这是国家利益使然，也是每一个具体的制造型企业提高企业竞争力的刚性需求。而大数据贯穿于企业生产到消费者消费流程，其中各个环节在互联网上沉淀的大数据可以交叉复用和服务于整个流程的各个角色，这意味着利用大数据响应和解决各个环节的“痛点”需求是整个行业不断发展和推进的必然。

在 2008 年金融危机中，阿里平台的交易记录预测了经济指数的下滑。2008 年年初，阿里巴巴平台上整个买家询盘数急剧下滑，预示了经济危机的来临。数以万计的中小制造商及时获得阿里巴巴的预警，为预防危机做好了准备。

百度拥有中国最大的消费者行为数据库，覆盖 95% 的中国网民，搜索市场占比达 87%。百度基于最真实的用户行为数据和多维度研究工具，帮助宝洁精准地定位了消费者的地域分布、兴趣爱好等信息。根据百度分析的结论，宝洁适时地调整了营销策略。

我国目前已经有十余座城市开展了数字医疗。病历、影像、远程医疗等都会产生大量的

数据并形成电子病历及健康档案。基于这些海量数据,医院能够精准地分析病人的体征、治疗费用和疗效数据,可避免过度或副作用较为明显的治疗,此外还可以利用这些数据进行实时计算机远程监护,对慢性病进行管理。

3) 后发优势使然,我国各行各业的IT基础建设不断成熟,并持续投入,为大数据应用提供了坚实的数据基础和基础设施保障。随着大数据概念不断得到认同(或许是市场发展现状倒逼“政产学研商用”各界认同),以“大数据”作为标签的大数据公司逐步独立发展,或是在原先业务的基础上延展起来,根据各个主体的不同优势和基础,各个公司主体着力于大数据产业链中的不同环节,而这个产业链也在不断成熟。

4) 政府应景发布的“大数据行动纲要”从政府层面为政府监管内的数据资源的开放共享设立时间表,这大范围降低了公司主体在进行大数据相关产业开发在数据收集(尤其是政府层面的数据收集)方面的成本,也为更广泛的数据应用提供了事实的数据基础支撑。

2015年9月5日,国务院发布了《国务院关于印发促进大数据发展行动纲要的通知》。文中指出,我国大数据发展的总体目标是立足我国国情和现实需要,推动大数据发展和应用在未来5~10年逐步实现以下目标:打造精准治理、多方协作的社会治理新模式;建立运行平稳、安全高效的经济运行新机制;构建以人为本、惠及全民的民生服务新体系;开启大众创业、万众创新的创新驱动新格局;培育高端智能、新兴繁荣的产业发展新生态。大数据发展的主要任务是:加快政府数据开放共享,推动资源整合,提升治理能力;推动产业创新发展,培育新兴业态,助力经济转型;强化安全保障,提高管理水平,促进健康发展。此外,提供足够的政策机制:完善组织实施机制;加快法规制度建设;健全市场发展机制;建立标准规范体系;加大财政金融支持;加强专业人才培养;促进国际交流合作。

综上所述,目前大数据的应用环境可以概括为:应用需求迫切、应用场景丰富、有政策法规支撑、有IT基础设施支撑、多边普遍认同。显然,所有这些都有助于大数据的广泛应用和推进。

3.2.3 技术环境

大数据这一概念得到“政产学研商用”各界的普遍关注和热议,各界的投入和既有的发展现状事实为大数据应用推进营造了良好的技术支撑氛围,具体而言(不限于):

1) 作为理性的研究者,为了响应大数据的需求和挑战,专门提出“数据科学”这一新兴学科领域(详见9.3节)以响应大数据涉及的理论及技术挑战。同时,围绕大数据技术流涉及的数据采集、数据组织与存取、数据建模与分析、系统开发和运维等各个方面的研究和教学也在持续推进中,为大数据产业的发展提供了坚实的技术支撑和人才支撑。

2) 基于开源精神的开源社区的不断发展,为大数据开源项目的丰富和发展提供了良好的平台和氛围。事实上已经不断得到广泛认同且有实质价值的大数据开源社区已经为大数据应用的不断推进提供了事实的基础(具体参见3.3.2节)。

3) 随着大数据产业链的不断成熟和明晰,处于不同角色及环节的大数据公司逐渐成长起来并产生了事实的领袖企业,这些领袖企业的典型产品为大数据应用的推进提供了实质的示范和借鉴。随着大数据产业的不断发展,大数据生态环境越来越稳定,从数据的产生到数据的应用,都将有一整套完整的技术作为支撑。

4) 特别值得一提的是,云计算的发展和推进为大数据应用的部署和运维提供了事实的基础设施保障。云计算作为计算资源的底层,支撑着上层的大数据应用(采集、存储、分析、运维)并将大数据的应用能力以云服务的方式提供给目标用户。根据目标用户和应用领域的特点,又分为公有云服务、私有云服务和混合云服务。

5) 在大数据场景下,“跨界融合、合作共赢”已经成为各界普遍认同的价值观,这意味着:在大数据项目研发和试错的过程中,通过跨界合作获得资源互补是大数据项目进程过程中的主旋律,这进一步加强了多边的合作,促进了集体智慧的涌现。

越来越多的传统金融交易和服务因互联网技术得以升级和替代,最直接的表现是支付方式的不断创新,电子支付系统不断完善并仍在继续发展。特别是以第三方支付为突破口,互联网企业利用网络平台和用户数据,在网络小额信贷等金融领域开始有所作为,在服务和技术等方面取得突破,对传统金融体系形成了挑战。我们所悉知的电商、O2O模式、众筹以及民营银行等都是融合经济发展进程中的产物。同时,得益于互联网的快速发展,金融服务正向着融合经济形态转变。客户数据是互联网支付公司的优势,跨界的融合为这些第三方支付公司带来了显而易见的好处。这些支付公司通过提供支付服务,积累了海量的零售客户资料和交易行为信息,且支付数据集中度高,通常横跨多个行业。通过运用大数据和云计算分析系统将用户行为模型化,可以形成资金流和信息流的交易数据闭环,这也就为他们提供更多的增值服务创造了可能。

综上所述,目前大数据的技术环境可以概括为:技术预研及人才储备持续进行中、基础设施建设逐渐完备、产业化分工逐步明晰、有可借鉴的示范应用。显然,所有这些都是有助于大数据的广泛应用和推进的。

3.3 大数据产业地图

3.3.1 大数据产业地图由来

由马特·图尔克(Matt Turck)主持编写的“大数据产业地图”(也称“大数据行业生态图谱”)是从风险投资者的角度对大数据的发展状况进行的陈述、研判和趋势解读,被业界(包括投资界和大数据产业界)誉为“大数据创业投资的清明上河图”,截至2014年5月,“大数据产业地图”已经发布到3.0版。

马特·图尔克曾任Bloomberg Ventures常务董事,2013年3月离职后,以合伙人身份加盟

FirstMark Capital, “大数据产业地图”是其在 Bloomberg Ventures 任职期间开始编制并持续维护版本改进的。

Bloomberg (彭博) 是 1981 年注册设在纽约的财务软件、数据和媒体公司, 经过二十余年的发展, 彭博已经成为全球最大的财经资讯服务提供商, 通过其强大的信息、专家和咨询网络为全球重要的决策制定者带来关键优势。彭博的优势在于通过创新的技术来快速、精准地传递数据、资讯和分析工具, 核心产品是彭博专业服务、彭博企业解决方案、彭博财经新闻、彭博新能源财经服务。Bloomberg Ventures (孵化器) 负责投资一些能够改善公司产品的项目, 但是 2013 年 3 月联合创始人马特·图尔克离任后几乎停滞, 现设立的 Bloomberg Beta 做类似的事情。

FirstMark 是 2008 年注册设在纽约的风险投资公司, 专注于风险资本交易, 关注的领域包括数据和分析、新兴的互联网和广告、云基础设施软件、垂直应用和解决方案。(摘自 wiki)

大数据产业地图尝试将现有的大数据产业在大方向上进行了一个较为明确的六大分类, 分别是:

- 1) 大数据数据源类: 匹配和响应大数据应用中的第一个问题“数据在哪里”。
- 2) 大数据分析类: 匹配和响应大数据应用中的第二个问题“如何使用数据”。
- 3) 大数据基础设施类: 匹配和响应大数据应用中的第三个问题“如何部署应用”。
- 4) 大数据应用类: 匹配和响应大数据应用中的第四个问题“应用在哪里”。
- 5) 跨基础设施分析类: 传统 IT 巨头为响应和拥抱大数据机遇而进行的业务延伸。
- 6) 开源项目类: 开源社区在大数据相关技术预研及应用延拓方面的工作。

大数据产业地图同时将这六大类当中数百个创业公司和 IT 厂商根据产品和商品模式划分成不同类别, 这种具有条理的划分方式有助于我们更好地梳理和了解大数据市场的发展。

3.3.2 大数据产业地图明细

大数据产业地图关于大数据产业六大类别的划分模式事实上是从三个维度描述了各个角色群体(大数据创新型企业、传统 IT 巨头、开源社区)是如何响应和拥抱大数据机遇的。以下详细介绍各个类别划分意义下的具体作为。

(1) 大数据数据源类

数据是大数据应用的基础, 数据的采集与整合是所有大数据项目的第一步工作, 互联网、因特网、万维网、物联网、移动互联网的迅猛发展以及在不同应用领域的渗透, 已经成为大数据事实的数据源, 所有这些数据或者以某个利益主体的私有财产的形式存放在私有服务器中, 或者以互联网开源的形式存放在互联网中(事实上, 这些数据物理上也存放在利益主体的服务器中), 这意味着, 任何一个大数据项目的开展都要从这些数据源中采集和整合相关数据, 显然, 这对于大数据项目的建设者而言, 都是不可回避的巨额投资。同时, 各个大数据平台自行收集和整合数据的做法也不利于数据的交叉复用和可扩展价值的彰显(具体见第 11 章)。出于行业分工的角度, 衍生出专门从事数据采集与整合的业务模式。根据数据源的不同以及应用模式的不同, 这类的企业又可细分为三种形态。

1) 数据市场: 将收集到的数据通过一个公共平台提供给数据的需求方, 这样的公共平台就能起到一个数据市场的作用。该平台承载着数据收集 (数据提供者向平台输入数据)、数据发布 (平台向数据使用者输出数据)、数据获益分配 (将数据使用者支付的费用以获益的方式分配给数据提供者和平台等相关角色) 等围绕数据的相关业务。

Data Market 就是一个具有“数据市场”意味的门户网站, 该网站通过三大核心业务 (功能) 实现数据的收集、整理和发布, 分别是: 数据传送引擎 (负责市场调研公司及其他信息发布者等数据拥有者或供应商更有效地提供数据)、数据中心 (采集和整合各个数据源数据)、开放数据服务 (帮助目标用户以统一的形式查看、下载和共享数据)。

2) 数据收集: 主动收集和整合某类 (些) 数据源数据, 然后将这些数据销售至潜在的目标用户群。这些数据源包括互联网数据和存放在私有利益主体服务器中的数据 (通过某些商务手段介入)。

随着互联网社交平台的逐渐普及, 在虚拟社交平台中存放着大量反映个体偏好、习惯、消费能力和倾向的数据以及人与人之间的关系的数据, 而且, 这些数据的活跃度巨大, 每天都有数亿级的用户在利用社交平台与他人进行交互, 其中产生的数据量非常惊人, 这类数据由于自身的特点, 使得它可以在如下方面有所作为: ①精准营销。利用社交媒体个性化数据在对的时间向对的人推销对的商品。②客户关系管理。在社交媒体上建立用户与品牌的互动关系。③社交媒体监测。媒体监测逐渐成为企业市场行为中不可分割的一部分。④科学研究。社交数据的爆炸为人工智能等领域提供了前所未有的机遇。因此, 有很多的数据提供商专门从事社交数据的收集与整合, GNIP 是其中最为著名的企业之一。GNIP 自 2010 年起便协同 Twitter 进行数据分析及销售, 其数据营销对象多为企业用户, 除 Twitter 以外, GNIP 公司也同时为 Facebook、Tumblr 及谷歌等多家 IT 公司提供相应的数据分析支持。鉴于 GNIP 公司向来以“丰富的数据资源及优质的分析技术”著称, 同时又一直和 Twitter 保持了良好的合作关系, 选择借 GNIP 之手完善自己的数据营销业务, 实为 Twitter 的明智之举, 因此在 2014 年 4 月 15 日, Twitter 宣布收购 GNIP。

3) 个人数据: 随着各类传感器以及各类穿戴设备的广泛使用, 探测个人各类健康数据已经变得非常便捷, 这类数据使得每个人都是一个独特的数据源, 对这种个人数据加以充分的利用能够有效地建立起个人健康模型, 从而为相关的目标应用提供辅助决策支撑。

BASIS 是世界上最先进的健康跟踪设备, 主要功能包括监测心跳规律、体表温度、卡路里消耗、单位面积出汗量, 这类数据因为和人的各项指标直接挂钩, 因此可以用来让用户了解自己的运动情况及健康状况, 利用这些数据来调节自身的生活习惯。在 2014 年, BASIS 公司的健康追踪智能穿戴产品约占 7% 的市场份额, 同年, 英特尔最终以 1 亿美元的价格收购了 BASIS 公司。

(2) 大数据分析类

数据分析是大数据的核心。在大数据场景下, 所谓数据分析指的是利用既有的领域知识

(经验) 或者利用某种手法从海量数据中挖掘出的规律模型(知识)操纵既有的数据,形成数据驱动的辅助决策。由于大数据的一个特点是“先有数据然后有模式”,这意味着大多情况下,领域知识往往不够完备,需要更多地仰仗从数据中发现知识和规律,从而体现出数据的价值。或许是因为大数据本身的复杂性使然,数据分析师的角色边界逐渐变得模糊,也就是说,数据分析师往往不再仅局限于数据分析本身,而会与前端的数据采集、领域分析及后端的系统实现、系统运维直接耦合。

1) 传统的数据分析师会按照既有的目标需求和数据进行目标导向的算法设计与预研,而在大数据场景下,数据分析师往往会被(甲方)问及:用这些数据你还能做什么?你还需要什么样的数据?这些数据在哪里?你知道如何获得吗?这就意味着,大数据场景下的数据分析师需要在(某)目标驱动下设计面向(甲方)利益最大化的整体解决方案,包括数据获取、数据存取、数据分析和计算架构、运维架构等。

2) 传统的数据分析师往往不需要顾及数据可视化的内容(传统的思路一般会认为这是美工或者人机交互的内容),而在大数据场景下,数据可视化的需求被刻意提高,这不仅仅是将复杂分析结果以良好的用户体验方式提供,还包括使整个数据分析的过程能够允许用户以可视化的方式回溯(二次研判)以及允许用户以可视化的方式进行数据建模,这并不意味着数据分析的职位完全取代美工、人机交互设计师,而是说在大数据场景下,数据分析师的团队中应该包含美工、人机交互设计师的角色。

3) 传统的数据分析师往往是以“一事一议”的方式进行工作,即根据给定目标设计算法然后交付,而在大数据场景下,如何充分复用和最大化数据分析师的分析能力被提到很重要的地位,产生这一现象的原因可能是:大数据的价值在于应用,而对于应用者(包括最终用户或者第三方开发商)而言,本来不用纠结于算法的设计与实现,而应该是思考如何根据目标导向,选择适当的算法进行系统的设计与实现。服务组合和服务计算为这一问题提供了一个很好的解决方案:将分析(计算)以服务的形式提供给第三方,就好像将数据(存取访问)以服务的形式提供给第三方一样,而这不仅是技术上的一种演变,也是很多创新应用模式和商业模式的基础。

在马特·图尔克的大数据产业地图中,大数据分析行业主要分成如下几种细分领域:分析解决方案、数据可视化、统计计算、社交媒体、舆情分析、分析服务以及IT分析等,此处不再赘述。

(3) 大数据基础设施类

发自奥巴马政府的大数据倡议的“从大数据中发现知识和洞见”的论调无疑引起了各方普遍关注和热捧,并逐渐上升为国家战略的重要渊源,为了实现这一目标,其技术流程无非是“数据采集→数据存储→数据分析→系统实现→系统运维”。如前所述,数据采集关注的是如何从(潜在)数据源获得数据;数据分析关注的是如何从数据中发现价值(知识和洞见),而支撑整个系统有效开发、部署和运维的基础是系统架构、数据存储、组织与管理与运维监管。

1) 作为分布式系统基础架构, Hadoop 无疑是市场占有率和认知度最高的一个, 其最核心的设计就是 HDFS 和 MapReduce, 前者为海量的数据提供了存储, 后者为海量的数据提供了计算。现如今, 企业和大型机构在寻求棘手的大数据问题的解决方案时, 往往会使用开源软件基础架构 Hadoop 的服务。许多公司都推出了各自版本的 Hadoop, 也有一些公司则围绕 Hadoop 开发产品。

典型的基于 Hadoop 的产品案例包括 Cloudera、Hadapt、MapR 等。

①Cloudera 由来自 Facebook、谷歌和雅虎的前工程师杰夫·哈默巴切 (Jeff Hammerbacher)、克里斯托弗·比塞格利亚 (Christophe Bisciglia)、埃姆·阿瓦达拉 (Amr Awadallah) 以及甲骨文前高管迈克·奥尔森 (Mike Olson) 在 2008 年创建。时至今日, Cloudera 已从一家当初默默无闻的创业公司, 发展成为企业在应对数据挑战时不得不依赖的公司。详细内容参见其官网 <http://www.cloudera.com>。

②Hadapt 是个自适应分析平台, 通过可以自定义分析的 Hadapt Development Kit (HDK) 和 Tableau 软件集成, 为 Apache Hadoop 开源项目带来了 SQL 实现, Hadapt 允许进行基于 SQL 大数据集的交互分析, Hadapt 2.0 成为了 Hadoop 工业上第一个交互式应用程序。详细内容参见其官网 <http://hadapt.com>。

③MapR 是 MapR Technologies Inc 的一个产品, 号称下一代 Hadoop, 使 Hadoop 变为一个速度更快、可靠性更高、更易于管理、使用更加方便的分布式计算服务和存储平台 (且是开源的), 同时性能也不断提高。MapR 号称不会出现 SPOF 单节点故障, 且被认为是与现有 HDFS 的 API 兼容, 因此非常容易替换原有的系统, 它能够为客户节约一半的硬件资源消耗, 使更多的组织能够利用海量数据分析的力量提高竞争优势。详细内容参见其官网 <http://mapr.com>。

2) 数据存储关注数据存在哪里。大数据应用底层离不开硬件存储部件, 因为大数据应用自身的特点使得其底层硬件的存储结构和其他存储结构有所区别, 也正因此, 涌现了一批采取不同技术和策略的存储型厂商。

Panasas 和 Nimble Storage 是两个典型的存储厂商:

Panasas 公司是一家为 Linux 集群提供可扩展网络存储方案的领导者, Panasas 的解典型特点包括: ①Panasas ActiveScale 存储集群将面向对象的分布式文件系统与灵巧的硬件相结合, 简化了 Linux 集群计算, 大大提高了性能和可管理性; ②Panasas 存储集群充分利用其单一的全局名字空间, 消除了烦琐的操作任务, 将管理成本降至最小; ③Panasas 系统提供了并行的数据路径和创纪录的 I/O, 使昂贵的集群投资获得最大的回报; ④Panasas 可以按 Linux 集群环境定制存储系统, 并能够无缝地集成到现有的数据中心架构中。

Nimble Storage 是以混合型存储闻名世界的存储厂商, Nimble Storage 以提供 SSD 与 HDD 混合存储阵列而闻名, 其间采用的新兴技术使得混合存储阵列既具备 SSD 的性能优势, 又具备 HDD 廉价的每 GB 成本优势, 在金融、政府、云计算、物联网等领域已经获得了很好的应用。

3) 数据组织与管理关注的是数据如何组织与存取。SQL、NoSQL、NewSQL 是大数据场景

下三类主要的数据库：SQL 凭借着自身独有的特点，例如获取持久化数据、对于事务的 ACID（原子性、一致性、隔离性、持久性）特性以及标准关系模型使得其在面对各种应用的时候都有着出色的表现，已经成为计算机文化的一部分；NoSQL 意为“不仅仅是 SQL”（Not Only SQL），是一项全新的数据库革命性运动，其诞生之初就是为了解决大规模数据集合多重数据种类带来的挑战，尤其是大数据应用难题，事实上在很早就有人提出，发展至 2009 年趋势越发高涨；NewSQL 是对各种新的可扩展、高性能数据库的简称，这类数据库不仅具有 NoSQL 对海量数据的存储管理能力，还保持了传统数据库支持 ACID 和 SQL 等特性。

目前 NewSQL 系统大致分三类：

①采取了不同的设计方法设计了全新的数据库平台，比如 Google Spanner、VoltDB、Clustrix、NuoDB（这类数据库工作在一个分布式集群的节点上，其中每个节点拥有一个数据子集，SQL 查询被分成查询片段发送给自己所在的数据节点上执行，这些数据库可以通过添加额外的节点来线性扩展）；NewSQL 还有一种策略是使用单一主节点的数据源但配备一组节点用来做事务处理，这些节点接到特定的 SQL 查询后，会把它所需的所有数据从主节点上取回来后执行 SQL 查询，再返回结果。

②设计高度优化的 SQL 存储引擎，这些系统提供了与 MySQL 相同的编程接口，但扩展性比内置的引擎 InnoDB 更好，比如 TokuDB、MemSQL。

③透明分片，这类系统提供了分片的中间件层，数据库自动分割在多个节点运行，比如 ScaleBase、dbShards、Scalearc。

大数据存储技术路线除了上面围绕 Hadoop 技术的相关产品还有采用 MPP（Massively Parallel Processing）架构的新型数据库集群，重点面向行业大数据，采用 Shared Nothing 架构，通过列存储、粗粒度索引等多项大数据处理技术，再结合 MPP 架构高效的分布式计算模式，完成对分析类应用的支撑，运行环境多为低成本 PC Server，具有高性能和高扩展性的特点，在企业分析类应用领域获得极其广泛的应用。

MPP 的一个典型例子就是 Teradata，它在一开始就使用 MPP 架构，以软硬一体机的产品方式提供给客户，其定位是高端客户的数据仓库和决策分析系统，Teradata 在全世界的客户只有几千个。在这个数据分析高端市场上，Teradata 一直是老大，在数据分析技术上 Oracle 和 IBM 无法与之抗衡。

4) 系统运维保障方面，安全问题被认为是其中一个最为关键的话题。在大数据时代下数据安全主要有如下 6 个方面的挑战：①大数据的巨大体量使得信息管理成本显著增加；②大数据的繁多类型使得信息有效性验证工作大大增加；③大数据的低密度价值分布使得安全防护边界有所扩展；④大数据的快速处理要求使得独立决策的比例显著降低；⑤大数据独特的导入方式使得攻防双方地位的不对等性大大降低；⑥大数据网络的相对开放性使得安全加固策略的复杂性有所降低。

Stormpath 是大数据安全方面具有代表性的产品,它提供了完整的用户管理和认证服务。借助这些服务,应用程序能够通过一个简单的 API 调用实现用户认证。账户注册、Email 验证、密码重置以及类似的功能都是内置的。通过 Stormpath,开发人员能够从安全问题解放出来,更加专心地解决应用程序功能需要。详细内容参见其官网 <http://www.stormpath.com>。

在 马特·图尔克的大数据产业地图中,大数据基础实施类行业主要分为 NoSQL 数据库、Hadoop 相关产品、NewSQL 数据库、MPP、管理监控等,此处不再赘述。

(4) 大数据应用类

落地应用是针对大数据这一个概念不断热炒的理性响应,没有明确应用导向、不能落地的大数据项目是典型的空中楼阁,3.4.2 节会详细研判大数据(技术)在各个应用场景的应用可能。在 马特·图尔克的大数据产业地图中,大数据应用类行业主要罗列和介绍了广告优化、出版工具、市场营销、行业应用、大数据应用服务提供商等若干种,前三者关注的是大数据的应用场景,而第四点关注的是企业的应用模式和商业模式。特别注意的是,这是 马特·图尔克从投资商的角度对大数据市场的描述和分析,并不是全集,此处不再赘述。

(5) 跨基础设施分析类

在 马特·图尔克的大数据产业地图中,为响应和迎合大数据的机遇与挑战,传统 IT 巨头(包括微软、IBM、Oracle 等)在既有产品线上进行拓展和延伸而展开的相关开发统一归并到“跨基础设施分析”类。

IT 行业是一个发展迅速的行业,在新的风口下,如果一家企业没有做出及时的调整,不能够快速地占领市场,那么就有可能被其他公司挤垮,历史上有太多这样的公司:曾经风光无限,在某一潮流档口顷刻倒塌。在大数据浪潮中,各种传统 IT 巨头如果不对这样的潮流做出及时的反应和改变,那么很可能就会被这样的潮流给吞没,最终被淘汰。出于响应挑战、占领市场并达到继续维系业界霸主地位的原始动机,几乎所有的 IT 巨头都在进行面向大数据市场的技术改良、产品改进。

以微软为例,微软把自己定位为可用性和大数据领域的领袖,最新版本的 SQL Server 纳入了大数据精神与 Hadoop 的提供商 Cloudera 合作,提供 Linux 版本的 SQL Server ODBC 驱动,开发 Hadoop 的连接器,跨入 NoSQL 领域。

Oracle 应景地推出的 Oracle 大数据机,Oracle 大数据机是一个硬、软件集成系统,融合了 Cloudera 的大数据产品及开源语言 R,期望为用户提供深入的大数据分析服务。同时,甲骨文公司还宣布推出了最新软件产品 Oracle Big Data Connectors,该产品可以帮助客户利用 Oracle 数据库 11g 轻松整合存储在 Hadoop 和 Oracle NoSQL 数据库中的数据。

SAP 研发的 HANA 数据库是一个软硬件结合体:软件方面,HANA 的内存数据库是其重要组成部分,包括数据库服务器、建模工具和客户端工具。HANA 的计算引擎是其核心,负责解析并处理对大量数据的各类 CRUDQ 操作,支持 SQL 和 MDX 语句、SAP 和 non-SAP 数据;硬件方面,SAP 和多个硬件厂商合作生产支持 HANA 的高性能服务器,包括 Dell R910、Fujitsu、

HP DL580、IBM x3850 等，以及和 Cisco 等公司合作。

(6) 开源项目类

开源 (Open Source)，即可开放源代码，用于描述那些源码可以被公众使用的软件，并且此软件的使用、修改和发行也不受许可证的限制，该术语（在软件领域）被非营利软件组织（美国的 Open Source Initiative 协会）注册为认证标记。

开源是一种文化，也是一种精神，简单来说，开源软件就是源代码开放的软件（事实上有一系列的属性特征，比如 Free Distribution、Source Code、Derived Works、Integrity of The Author's Source Code、No Discrimination Against Persons or Groups、No Discrimination Against Fields of Endeavor、Distribution of License、License Must Not Be Specific to a Product、License Must Not Restrict Other Software、License Must Not Restrict Other Software、License Must Be Technology-Neutral 等）。对普通用户来说，是否开源其实意义不是很大，不过对于商业用户来说，开源的意义就很大，比如：减少开发周期、降低开发成本、保障产品质量等。

随着大数据应用的逐步推进以及相关技术的不断成熟，开源作为底层技术授权解决方案的最大贡献者的优势越来越明显。一个事实是，现如今，从小型初创企业到行业巨头，各种规模的供应商都在使用开源来建设和开发大数据项目，借助开源以及其他相关技术，新兴公司甚至在很多方面都可以与大厂商抗衡。在大数据方面的开源工具包括四个领域，分别是：

1) 数据存储。比如 Apache Hadoop（其官网是 <http://hadoop.apache.org/>）、MongoDB（一种 NoSQL 数据库，其官网是 <http://www.mongodb.org/>）、Cassandra（一种 NoSQL 数据库，其官网是 <http://cassandra.apache.org/>）、Hbase（一种 NoSQL 数据库，其官网是 <http://hbase.apache.org/>）、MySQL（一种 SQL 数据库，其官网是 <http://www.mysql.com/>）、MariaDB（一种 SQL 数据库，其官网是：<https://mariadb.com/>）、PostgreSQL（一种 SQL 数据库，其官网是 <http://www.postgresql.org/>）、TokudB（一种 SQL 数据库，其官网是 <http://tokutek.com/>）、Apache Drill（一种海量数据交互式查询引擎，其官网是 <http://drill.apache.org/>）、Apache Sqoop（一个用于将 Hadoop 和任意关系型数据库中的数据进行数据交换的工具，其官网是 <http://sqoop.apache.org/>）。

2) 开发平台。比如 Apache Hadoop 平台（其官网是 <http://hadoop.apache.org/>）、Apache Lucene（一种全文检索平台，其官网是 <http://lucene.apache.org/core/>）和 Solr 云平台（其官网是 <http://lucene.apache.org/solr/>）、OpenStack（一种云平台，其官网是 <http://www.openstack.org/>）、Red Hat（搭载 Hadoop 服务标准的 Linux 平台，其官网是 <http://community.redhat.com/>）、REEF（微软的 Hadoop 开发者平台，其官网是 <http://www.openreefs.com/>）、Apache Storm（集成计算和存储的开发平台，其官网是 <http://storm.apache.org/>）、Apache Spark（集成计算和存储的开发平台，其官网是 <http://spark.apache.org/>）。

3) 开发工具和集成。比如 Apache Mahout（机器学习编程语言，其官网是 <http://mahout.apache.org/>）、Python（脚本语言，其官网是 <http://www.python.org/>）和 R 语言（预

测分析编程语言,其官网是 <http://www.r-project.org/>)。

4) 分析和报告工具。比如 Gephi (一款开源免费跨平台基于 JVM 的复杂网络分析软件,官网 <http://gephi.github.io/>)、JasperSoft (报告和分析服务器,其官网是 <http://community.jaspersoft.com/>)、Pentaho (数据集成和业务分析,其官网是 <http://community.pentaho.com/>)、Splunk (IT 分析平台, <http://www.splunk.com/>)、Talend (大数据集成平台,其官网是 <http://www.talend.com/>)。

大数据开源社区对大数据技术的发展和推进有着非常积极的作用,这些开源社区(尤其是官网)不仅提供了相关阅读文档,有助于更简便地学习相关的工具,还提供了源码下载,对从事大数据相关技术的学习者大有裨益。在 马特·图尔克的大数据产业地图中,大数据开源项目被细分为框架、查询/数据流、数据访问、协作/工作流、实时、统计工具、机器学习、云部署等类别,此处不再赘述。

3.3.3 大数据产业地图意义

在大数据时代,马特·图尔克以图形化方式描述的大数据产业地图得到了多边人士的广泛认可,并成为投行甚至从事大数据相关研究和应用的风向标,究其本质,或许是因为该产业地图从产业结构梳理的角度明晰了大数据产业的方方面面,为相关人士期望在大数据领域的行动和决策提供了一个研判基础,应用提示在于(不限于):

1) 大数据产业地图是马特·图尔克从投资者的角度描述了(每个版本发布时)大数据产业的现状(静态数据),并根据不同版本的描述变化(动态数据)分析大数据产业(各个环节)的演化趋势。

2) 显然,大数据产业地图描述的产业现状对国家在大数据方向的产业政策制定具有务实的参考意义。

3) 正所谓“他山之石,可以攻玉”,来自投资者的研判白皮书对于大数据“淘金客”而言,其实质的参考价值或许比来自国家意志的导向性文件更加直接和务实。至少,通过大数据产业地图,可以清晰地知道潜在的竞争对手或业界标杆正在做什么。

4) 对于从事大数据关键技术预研的研究者而言,也能通过大数据产业地图获悉大数据应用场景,这对于应用驱动的应用研究者而言,价值颇大。同时,各个技术型公司的所作所为(特别是开源社区发布的在大数据方面的工作和进展)也能够为关键技术及理论研究的工作者提供可借鉴的思路。

5) 对于期望学习和了解大数据(这一概念及产业)的(独立)第三方而言,大数据产业地图也能够对市场现状、产业现状、技术研究现状等方面给出相对完备的综述和索引,这对于有兴趣或者有志于大数据的群体而言,大有裨益。

综上所述,马特·图尔克的大数据产业地图具有一定的地位和指导意义,但是应当注意到:

1) 马特·图尔克是从投资者的角度给出大数据产业(行业)的划分,但是划分的方式并不是唯一的,可以从技术层面划分(数据→存储→分析→架构→应用),也可以从商业模式

(软件、咨询、服务等)进行划分。不管如何划分,许多公司业务(模式)是跨类别的。

2) 大数据产业地图罗列的公司并不是全集,更多的未容纳进来,特别值得一提的是,马特·图尔克产业地图基本没有涉及中国的相关企业(或许是马特·图尔克手边并没有中国企业的数据,或者其从投资策略的角度而言没有考虑中国市场等),但是中国有关企业在相关大数据领域的探索却并不落后(比如百度的“百度大数据引擎”、腾讯的“腾讯云分析”、阿里的“阿里云”等)。

3) 除了马特·图尔克发布的大数据产业地图外,很多的咨询机构或公司也会定期发布大数据行业的发展白皮书,这些白皮书对于了解大数据的发展现状同样重要。而我国不同级别的产业协会、学会或专委会也会定期(一般以自然年度进行)发布中国大数据发展趋势白皮书,对于了解有中国特色的大数据现状,同样大有裨益。

3.4 大数据应用提示

3.4.1 大数据中文解析及提示

在大数据这一概念还没有像现在这样得到广泛认同的时候,针对海量数据的分析和处理就一直在进行,特别是在一些信息化发展相对较快的行业(比如天文、气象、金融等),已经有很多成熟的应用。以金融行业的数据分析为例,早在1980年,劳伦斯·克莱因(Lawrence Robert Klein)就因建立了经济体制的数学模型而获得诺贝尔经济学奖;而特里夫·哈维默(Trygve Haavelmo)则将经济与计算牢牢地结合在一起,奠定了计量经济学的基础,也因此获得了1989年的诺贝尔奖。随着社会各行各业信息化程度不断加深,海量数据的涌现为数据分析技术提供了更为广阔的舞台,兼具数据和计算于一体的“大数据”这一概念事实是在长期技术积累和经验积淀的基础上“横空出世”的。

有一种理解可以用来很好地分析和遴选大数据的应用场景,即大数据=“大”+“数”+“据”,其中文解释及应用提示如下。

(1) “大”

《说文解字》中提及“大,天大、地大、人亦大,故大象人形”,意为“面积、体积、容量、数量、强度、力量超过一般或超过所比较的对象,与‘小’相对”或者专门指“规模广、程度深……”,这对于大数据应用的提示是:

1) 大数据需要追求数据(数)的更广、更深,唯有如此才能使得利用既有的先验知识或者数据挖掘获得的知识处理更多的事务。互联网的持续发展及应用的拓展和推进为大数据的采集和整合打下了坚实的基础。

2) 大数据应该追求更完备、更应景的知识,唯有如此才有可能使得利用知识处理和分析数据不会出错。互联网的持续发展及应用的拓展和推进为“利用知识处理数据”获得的价值得到更广范围的渗透。

(2) “数”

《说文解字》中提及“数，计也”，意为“计算”，后来演变为“表示、划分或计算出来的量”，如“五陵年少争缠头，一曲红绡不知数”（白居易《琵琶行》），鉴于计算机环境下，所有的数据最后都格式化为可计算的“数”，因此：

1) 大数据环境下，必须将所有数据格式化为计算机可以计算的“数”，否则无法充分利用计算机的计算能力。

2) 大数据环境下，所有的“数”必须被计算（和使用），否则没有任何意义，这种计算和使用包括利用既有的（先验）知识分析和处理、利用专门的手段从数据中发现知识，然后利用这种知识进行后续的分析 and 处理。

(3) “据”

《说文解字》中提及“据，杖持也”，意为“凭依、依据”，可以理解为处理事务的依据或者知识。在大数据应用场景下，应该理解为利用这种依据或者知识分析和处理所采集到的“数据”（而在计算机环境下，所有的数据最后都格式化为可计算的“数”）。潜台词是：

1) 在有先验知识的情况下，可以利用知识直接分析和处理数据，这意味着在大数据应用场景下，如果有完备的领域知识（先验知识），可以直接操作数据，而无须进行数据建模。

2) 在没有知识的情况下，需要设法获得知识，而在大数据环境下，知识的来源唯有一大堆数据，需要通过数据建模从数据中挖掘出知识来。鉴于大多数场景下，我们往往并不具备（明确）具体领域的知识，因此“从数据中发现知识和洞见”被认为是大数据应用中的重要环节，也是奥巴马政府提出“大数据研究和发展倡议”时的目标。

基于上述分析，在进行大数据应用场景的遴选和研判的时候，可以遵循的一些思路或原则包括（不限于）：

1) 该应用场景应该具备良好的甚至坚实的 IT 基础，一般而言，应该是互联网化改造和建设成熟的，否则谈大数据等于是空穴来风，从无到有地建设一个系统是信息化的工作，而不是大数据的事。当然，针对这个应用场景有甲方愿意在短期内投资也是必须考虑的一个指标，这就与这个应用场景的价值评估相关，而这一点，与大数据本身或许无关。

2) 该应用场景应该具备足够规模的数据（包括数据的活度、厚度及混杂度等），并且这些数据能够便于数字化和存储（否则计算机无法处理），当然，这也与此应用场景的 IT 基础建设直接相关。另外一个需要考虑的问题是：现有技术是否能够从容地理解和分析这些数据，如果还没有的话，需要进行这方面的理论及技术攻关，而不能匆忙上马大数据项目。

3) 该应用场景是否具备足够的先验领域知识或者能否从既有的数据中挖掘出有效的知识（模型）也应该是一个重要的考察指标。前者关注的是人们对相应的领域是否有足够的认知度，如果没有的话，需要从研究的角度去进行探索（大数据应用于某个领域的研究也是一个大数据的发力点，但这个与具体的应用往往无关）；后者关注的是建模技术及分析技术的成熟度，这会直接影响到大数据项目的建设成本与进度。

3.4.2 大数据应用场景及策略

在大数据概念热炒的当下，各个利益主体（包括政府、企业和期望在大数据市场一展身手的个人）在大数据项目建设方面给以足够的重视和投入，这显然是利好的。但不可否认的是，许多大数据项目的规划是盲目跟风的，或者说，许多大数据项目的建设本身是不具备上述大数据应用场景遴选精神的，大多问题体现在（不限于）：

1) 该应用领域的 IT 建设基础薄弱，从无到有地构建大数据项目如同空中楼阁，看起来很美，但是没有基础，换句话说，大数据项目应该首先落地在信息化建设成熟的场景。

2) 该应用领域的价值期望过低，难以获得更多的经费支撑，而大数据项目建设本身首先是个持续“烧钱”的过程，投资收益的盈亏平衡点往往战线拉得很长，换句话说，大数据项目应该首先落地在有持续经费投资的场景。

3) 该应用领域的数据活度过低，难以产生“数据→价值”的涌现，换句话说，大数据项目应该首先落地在（用户交互）频度大、数据（增量）活度大的场景。

4) 该应用领域的的数据收集困难太大，或者是投资成本太高，或者政策法律瓶颈使得根本无从收集相关数据，显然，没有相当规模的数据支撑，谈大数据项目只是空穴来风，换句话说，大数据项目应该首先落地在数据收集难度可控的场景。

5) 该应用领域的知识储备过低，缺乏对这个领域最起码的认知，尽管大数据的价值是从数据中发现知识和洞见，但这应该建立在对这个领域具备最原始的公理系统（这往往是容易被忽视的一个原则）的基础上，否则大数据项目的开展无从下手，换句话说，大数据项目应该首先落地在有（一定）认知度（或者有公认的方法论体系）的场景。

6) 该应用领域的大数据项目建设缺乏必要的技术支撑和人才储备支撑，前者涉及的是技术选型问题，后者涉及的是项目的运维。如果该领域的大数据项目涉及无法解决的技术瓶颈，这个项目会因为依赖条件的无法匹配而无限期搁置（延误）；如果此大数据项目没有配备持续的人力加以运维，这个项目也会因为不能持续运转而被搁置。换句话说，大数据项目应该首先落地在各方面保障完备的场景。

通俗地说：“有价值、有基础、能实现”是大数据项目是否能够成功落地的必要条件。有价值是指该应用场景的价值能够得到多边共识并有投资方愿意（持续）投资；有基础是指该应用领域应该具备较好的 IT 基础、较好的数据基础以及较好的人力（才）储备基础；能实现是指项目建设没有不可解决的技术瓶颈、必须有充分的方法论体系作为支撑、必须有运维团队保障大数据平台的落地运维等。

观摩一些领袖企业在大数据时代的产品规划也可以发现类似的价值观。以业界航母 IBM 为例，在拥抱大数据的商业实践方面，IBM 产品线的几个着力点在于：

1) 利用大数据探索实现信息库的充实。

此方面针对的应用场景及需求是客户服务、保险、汽车、维修、医药等行业需要储备规模巨大的知识库，而庞大繁杂的解答手册和知识系统会造成重复查询，导致系统延迟和成本

上升。

2) 利用增强 360 度全方位客户视图实现客户交互改进。

此方面针对的应用场景及需求是电信、零售、旅游、金融服务和汽车等行业将“快速抓取客户信息从而了解客户需求”列为首要任务。

3) 利用运营分析实现运营优化。

此方面针对的应用场景及需求是制造、能源、公共事业、电信、旅行和运输等行业需要时刻关注突发事件、通过监控提升运营效率并预测潜在风险。

4) 利用数据仓库扩充实现 IT 效率和规模效益提升。

此方面针对的应用场景及需求是企业需要增强现有数据仓库基础架构,实现大数据传输、低时延和查询的需求,确保有效利用预测分析和商业智能实现性能的扩展。

5) 利用安全性和智能扩展实现犯罪防范等。

此方面针对的应用场景及需求是政府、保险等行业亟待利用大数据技术补充和加强传统的安全解决方案。

3.4.3 大数据陷阱及应用提示

大数据为我们的生活、工作与思维带来深刻的变革,在享受大数据所带来的社会进步的同时,我们也应当理性地注意到大数据自身所存在的问题与陷阱,或许这些问题和陷阱并非是大数据本身的原罪,但恰恰是在进行大数据应用的过程中,作为理性慎思的人们需要考虑和回避的,比如(不限于):数据封闭问题、数据断裂问题、数据隐私问题、数据歧视问题、数据独裁问题、数据垄断问题等。

(1) 数据封闭问题

随着信息化手段在人们工作、生活及娱乐等各个环节的应用不断深入和渗透,几乎反映实体对象(人、物或者事件)的数据都被记录和存储了下来,通过对这些数据的有效分析,应该能够充分发挥数据的价值。但是摆在数据分析师面前的往往是“有所有数据但什么数据都欠缺”的尴尬,一个重要的原因在于:所有这些数据都散布在不同的业务平台中,而这些业务平台往往归属彼此利益独立的利益主体,比如腾讯存储了人们在 QQ 和微信上的言论(关系)数据、阿里记录了人们在淘宝上的购物数据、百度记录了人们利用搜索引擎的搜索数据、移动运营商记录了人们的日常通信数据……这些数据封闭(数据被封闭在不同的业务系统,也是一种“数据孤岛”)的现状使得大数据分析师无法获得多样化的数据,从而阻碍数据价值的实现。当然通过一定的商业模式,大数据项目建设者可以通过购买和交换的方式获得这些数据,但无疑增加了大数据项目的建设成本,也使数据应用的灵活性大为降低。在大数据产业链中,出现了以“数据集市”为标签的商业模式(第三方作为中间商收集、整合和共享数据,在数据层面打破现实世界的界限)或许会是一个解决数据封闭的途径。

(2) 数据断裂问题

数据断裂问题指的是数据缺乏结构化、物理实体与虚拟实体或者虚拟实体与虚拟实体之

间缺乏有效的映射,这就使得多源数据的整合成为极为棘手的问题,而上述的“数据封闭”则进一步强化了这个问题。封闭数据使我们无法看到多样化的数据,断裂数据则使数据缺乏结构化。来自 IDC 的报告显示,2012 年全球数字信息中 90% 的数据都是视频、声音和图像文件这样的非结构化数据,缺乏结构化本身是可以借助新技术解决的问题(这也吸引了大量的研究精力),正因为如此才使这个问题变得棘手:对新技术的过分追逐,一方面会使得数据本身的真实性、完整性遭到破坏,另一方面会使对数据背后的人和生活意义的分析得不到充分重视。总之,数据封闭与数据断裂问题,阻碍了全数据模式的实现,企业将数据视为财富,但却形成“数据孤岛”的境地,不利于全社会数据的共享。

(3) 数据隐私问题

数据隐私问题是大数据时代的一个重要问题,众所周知,“数据为王”是大数据时代的共识,基于这种共识而产生的“数据淘金狂潮”激励商家使用各种手段采集和整合个人数据;而作为一个普通消费者的个体,往往也是在有意地享受商家服务的同时将隐私无意识地拱手相让;而政府从国家管理的角度也在有意地收集和记录民众隐私,这本无可厚非,但如果这些数据被不法分子非法获得并被加以利用,不管最终目标是什么,隐私总归是泄露了;而国与国之间的大国博弈,也使得隐私数据在安全博弈的场景下被有意收集和整合。

(4) 数据歧视问题

数据歧视问题指的是过于依赖大数据分析结果而产生的对事件本质的误判。一个来自于科幻电影《少数派报告》的极端例子是:假如人类数据分析能力已强大到可预测人类个体的犯罪行为何时发生,此时会有一个可怕的伦理问题,即人类会为即将可能发生的犯罪负责而不仅仅是对“已做”的负责,而这显然是有失法律精神的,也是否认了本该有的自由意志并伤害了人类尊严。当然,这未必是数据本身的错误,而是把大数据非理性地应用到它不适用的领域而导致的。

(5) 数据独裁问题

数据独裁问题指的是过于依赖和迷恋数据本身而忽略了事件的本质,这个问题其实一直存在,而并非大数据时代专有,比如政府使用绩效来评定一个官员的执政水平,这就使得官员过分注重绩效数字本身,而不是真正的“执政为民”;出于对时间和人力成本考虑,公司招聘时将学历较低或成绩较低者直接拒之门外,而不考虑这些不达标者的真实能力水平,这也造成了求职者对成绩和学位而非真实能力的执着。随着大数据的发展,人们会越来越想从大数据中掘金,最终导致一种对数据的盲目崇拜和认同,数据独裁会持续发酵。

(6) 数据垄断问题

数据垄断问题指的是具有垄断地位的大数据公司因为对数据(数据之于信息时代就如同燃料之于工业革命)具有垄断地位而成为信息时代的垄断企业,使得本该公平的竞争从一开始就处于不公平的状态。众所周知,企业的发言权取决于和最终用户的靠近程度,软件的价值度取决于所管理数据的规模和活性,市场的支配权取决于对数据的掌控力,而当谷歌拥有用户的一切数据或者亚马逊知晓顾客购物行为的一切时,其他新的公司根本无法与谷歌和

亚马逊竞争,在这种情况下,谷歌或者亚马逊就是因为对数据的垄断地位而成为(信息时代的)像铁路、电信一样的垄断公司。

大数据的技术流是:数据采集→数据存储→数据建模→应用开发→系统运维,上述提及的若干陷阱和问题(或许总结得还不够全面)会在不同的环节单独或者并行出现,这也意味着,在大数据项目整个生命周期内,我们都要审慎地考虑和响应。

“我们相信上帝,除了上帝,其他任何人都必须用数据说话。”这是现代经理人的信仰,也回响在硅谷的办公室、工厂和市政厅的门廊里。对大数据善加利用,这是极好的事情。但是一旦出现不合理利用,后果将不堪设想。

3.5 本章小结

大数据是数据本身以及数据采集的工具、平台和分析系统的总称,粗放地说,所谓“大数据”就是:“大”+“数”+“据”。

互联网作为一个普适互联的虚拟网络,它的迅猛发展,使得人们的工作、生活、娱乐等各类活动在一个更加开放、平等、互动、不断迭代和演化的平台上进行的同时,大量的数据因互联网而产生并沉淀在互联网中,可以说,没有互联网及互联网的持续发展,就不会有目前称之为“大数据时代”的来临。

信息化水平的不断提升,使人们得以用更丰富、更多元的手段去采集和收集更多种类、更多维度的数据,并因为有更便利、更快捷的存储使得人类能够保留比历史上任何一个时代更多、更广的数据,“万事万物数字化”已经成为事实,并成为“大数据”的产生基础,并因为“数据交叉复用化”这一策略的践行使“大数据”这一思维得到持续推进。

最为重要的是,作为这个地球上最为高等的生物,人类出于自身的存在或者更好存在的原始需求在人类发展过程中持续膨胀,特别是在信息化水平不断提升的当代,这种需求的膨胀越来越大,这或许才是大数据得以广泛认同并得到迅猛推进和发展的最原始动力。

总之,我们现在正沉浸在一个标签为“大数据”的时代,出于各边利益的诉求,“政产学研用”各界包括老百姓个体都对这个称之为“大数据”的东西表示出了极大的认同、热情、期待和投入。政府出于数据主权博弈及社会治理和发展的形势制定出台了若干引导性和倡议性政策、指南及规划,为大数据的发展提供了良好的发展空间,而且这几乎是各国政府都在做的事情;企业出于提供自身竞争力及获得更多收益的目标也在以各种行动拥抱大数据的各项机遇和挑战,大数据产业链处于不断成熟和良性发展的状态;大数据研究者也从理论技术预研的角度以百家争鸣的势头响应和迎接大数据带来的种种需求和难题;而处于战略转型和产业结构调整的大势及“中国制造2025”的风口,中国的企业(尤其是制造型企业)也在努力拥抱大数据。或许正是如此多边的集体发力,“大数据”正在从刚开始的媒体热炒的概念持续演变为媒体会继续热炒的落地应用。

本章简单地梳理了大数据生态环境,并以马特·图尔克制定的大数据产业地图分析了大

数据产业链的各个角色分工及分布,对大数据应用场景选型及研判给出了一些建议和原则。

狄更斯在《双城记》开篇提到:这是一个最好的时代,也是一个最坏的时代;这是明智的时代,这是愚昧的时代;这是信任的纪元,这是怀疑的纪元;这是光明的季节,这是黑暗的季节;这是希望的春日,这是失望的冬日;我们面前应有尽有,我们面前一无所有;我们都将直上天堂,我们都将直下地狱……

我们正沉浸其中的大数据时代,或许也正如此。

本章参考文献

- [1] Kumar R, Gupta N, Charu S, et al. Manage Big Data through NewSQL [C]. National Conference on Innovation in Wireless Communication and Networking Technology-2014, 2014.
- [2] MattTurck. The State of Big Data in 2014: a Chart [EB/OL]. <http://mattturck.com/2014/05/11/the-state-of-big-data-in-2014-a-chart/>, 2014.
- [3] Turck M, Zilis S. A Chart of The Big Data Ecosystem, Take 2 [EB/OL]. <http://mattturck.com/2012/10/15/achart-of-the-big-data-ecosystem-take-2/>. 2012.
- [4] 阿里研究院. 2014 年我国大数据发展分析报告 [EB/OL]. <http://www.aliresearch.com/blog/article/detail/id/19932.html>, 2014.
- [5] 涂子沛. “大数据”正在到来的数据革命 [J]. 求贤, 2013(12): 60-61.
- [6] 维克托·迈尔·舍恩伯格, 周涛. 大数据时代: 生活、工作与思维的大变革 [J]. 人力资源管理, 2013(3): 136.
- [7] 赵国栋, 易欢欢, 糜万军, 等. 大数据时代的历史机遇 [M]. 北京: 清华大学出版社, 2014.

第二篇 Part 2

技术及选型思路

- 第4章 大数据支撑技术
- 第5章 数据采集与整合
- 第6章 数据存储与管理
- 第7章 数据表示与理解
- 第8章 数据理解与建模
- 第9章 知识发现与应用

第4章 大数据存储技术 大数据与大数据技术课程教学案例解读与数据科学

二、

.....

.....

大数据支撑技术

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的戴恒宇、李红、蔡洋、高扬、陈嘉伟、唐驰等几位同学的协助，在此表示深深的谢意。

4.1 引言

据说是诺亚 600 岁生日的那天，天下洪水泛滥，诺亚按照上帝事先教导的方法和规格用歌斐木制造了诺亚方舟生存了下来……（摘自《圣经·创世纪》）

大约 4000 多年前，父氏社会末期的中国，洪水泛滥，大禹依靠“疏通”之法治水成功，并划天下为九州……（摘自《山海经·海内经》）

公元前 256 年，秦朝蜀郡太守李冰及其子率众以无坝引水之法修建都江堰，一改岷江洪水淫威下的成都平原为天府之国……

尽管考古材料还无法证实，但是仍然有很多理性的研究人员认为诺亚方舟和大禹治水的故事应该有其真实的历史原型，本文不对此历史的真伪进行点评。本章关注的是，神话传说和历史典籍都在向世人展示着古人的智慧：人类善于利用既有的经验和技术去迎接和响应新兴困难的挑战，本章开头的故事均是关于“治水”，在其他方面，比如计时、气象预测等也均能反映出人类的这一智慧。或许这也是人类能够通过自然选择而成为地球主宰的原因。因为进化论的研究者相信，会制造工具并使用工具是人类走向文明的一个重要基础。

本章关注的话题是：伴随着大数据时代的来临，大数据在带来发展机遇的同时也带来了若干挑战，如何有效迎接和响应这样的挑战是摆在学术界相关研究人员、工业界相关工作人员和最终应用方面前重中之重的难题。或许现有的技术还不能完全解决所有相关的问题，正如诺亚制造了方舟应付来自上帝对人类的惩罚、大禹发明了“疏通”之术治水划九州、李冰父子利用“无坝引水之法”成就了天府之国一样，当前的研究人员提出第四类研究范式并提出“数据科学”这一学科专门用于响应来自大数据的挑战。同时，我们也应当乐观地看到，

尽管大数据是一个新兴的概念，但是与处理大数据相关的技术已经发展了相当长的时间，正是这些技术的积累，才为大数据的发展提供了基础。

催生大数据技术迅速发展的是数据科学，数据科学（Data Science）是计算机科学与统计学的交叉学科。同时数据科学这一概念的出现也凸显了各界对大数据的兴趣和关注。数据科学是一门利用数据学习知识的学科，其目标是通过从数据中提取出有价值的部分来生产数据产品。它结合了诸多领域中的理论和技术，如应用数学、统计、模式识别、机器学习、数据可视化、数据仓库以及高性能计算等。数据科学已经在 IT、金融、医学、自动驾驶等领域得到广泛应用并对计算机视觉、信号处理、自然语言理解等多个计算机分支产生了重要的影响。

大数据支撑技术主要涵盖计算架构和算法。针对大数据的描述，有一种说法是：“应用为本、数据为王、算法为资、架构为帅”，其本意是说：

1) 大数据项目的开展必须是以某具体目标应用为根本导向的，否则大数据的价值无从谈起。

2) 大数据项目的开展一定是以数据为基础的，没有数据，大数据项目的开展就是典型的“无源之水、无本之木”，拥有数据是保持竞争力的根本。

3) 分析手法是大数据项目开展的核心技术保障，因为大数据的本质是通过数据分析的手法从数据中发现知识和洞见。

4) 大数据的 4V 特征给大数据分析带来了极大的挑战，算法分析的极高复杂度和海量分析吞吐率的矛盾必须依赖更快、更高效的计算架构来解决。

“资”和“帅”出自《墨子》，其原文是“德为才之帅，才为德之资。德器深厚，所就必大。德器浅薄，虽成亦小”。大意是说，德行是才学的统筹，才学是德行的资本。道德修养与才识度量都很好的人，他的成就必定很大。反之，即使有成就也是小乘。对比看算法和架构的关系，也是如此：没有良好的计算架构，算法即使很高明，由于计算复杂度很高，也很难应付和响应大数据的挑战。有了良好计算架构的支撑，高明的算法才会如虎添翼。而计算架构也必须要有好的算法运行其上，否则也是空有一个“皮囊”，无所大用。

计算机工作者的一般共识是“程序 = 算法 + 数据结构”。按照这样的思路类比，在大数据环境下，如果说大数据应用是与此对应的“程序”的话，数据科学就是大数据应用这个“程序”的“算法”，这个“算法”是大数据应用的核心，机器学习与数据挖掘算法是其中的关键支撑，为大数据应用提供理论支持。大数据应用的“数据结构”则更偏重（架构）技术方面，如当前非常流行的并行计算方法 MapReduce、Spark、Storm 等。鉴于单个计算机的计算能力不足以为大数据提供计算服务，因此计算机集群以及以集群为基础的计算方法快速发展起来。综上所述，仿照“程序 = 算法 + 数据结构”的说法，可以认为：“大数据应用 = 数据科学 + 计算架构”。

当前，以 BAT（百度、阿里巴巴和腾讯）为首的中国互联网公司在处理大数据上都是以

集群的方式来处理。其优点在于：

1) 高可伸缩性。服务器集群具有很强的可伸缩性，随着需求和负荷的增长，可以向集群系统添加更多的服务器。在此配置中，可以有多台服务器执行相同的应用和数据库操作。

2) 高可用性。高可用性是指在不需要操作者干预的情况下，防止系统发生故障或从故障中自动恢复的能力。通过把故障服务器上的应用程序转移到备份服务器上运行，集群系统能够把正常运行时间提高到99.9%以上，大大减少服务器和应用程序的停机时间。

3) 高可管理性。系统管理员可以从远程管理一个甚至一组集群，就好像在单机系统中一样。

一个大数据项目开展之初，需要回答的若干基本问题（至少）包括：

- 1) 数据从哪里获得？如何获得？
- 2) 数据将存在哪里？如何透明存取？
- 3) 如何分析、使用这些海量数据？
- 4) 面对海量数据，如何高效计算？
- 5) 如何运维？

本章尝试从大数据技术流程的角度阐述从数据获取到系统运维这一数据价值实现流程中涉及的环节以及各个环节涉及的相关技术和非技术因素，本章下面的结构安排如下：4.2节介绍大数据的处理流程，尝试描述大数据项目部署落地的项目全景以及各个环节面临的技术挑战，并给出大数据项目建设成效的评估方法；4.3节详细介绍大数据流程中各个环节涉及的关键技术，并对其他涉及的关键技术进行详细说明；4.4节介绍大数据项目在部署实施和系统运维中涉及的相关内容和应用提示；4.5节对本章进行小结。

4.2 大数据流程

图4-1扼要地给出了大数据流程的一般框架，如图所示，开展一个大数据项目涉及的内容是在数据计算架构和其他相关技术的保障下，从数据源采集数据并以便于数据管理和后续访问的方式进行数据存储，然后通过数据分析手法在数据理解的基础上进行数据建模和分析，通过数据建模和分析发现的知识和洞见，为目标应用提供数据支撑，同时开发面向目标应用的运维系统或服务平台。其中涉及的主要环节包括：

1) 数据采集。数据采集环节关注数据在哪里以及如何获得数据。

2) 数据存取。关注数据存在哪里以及如何透明存取。

3) 数据分析。关注如何分析和建模数据，从而发现数据背后的知识和洞见。

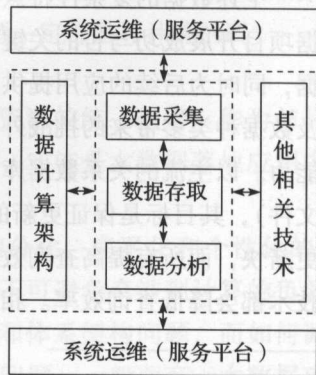


图4-1 大数据流程框架

- 4) 数据计算架构。关注如何高效计算以及系统的运维模式。
- 5) 服务平台。关注如何对接应用系统,为应用提供统一服务。
- 6) 其他相关技术。主要涉及保障上述流程有效、可信运转的技术和非技术影响因素。

4.2.1 显式挑战

该技术框架是一个普适的大数据流程框架,任何一个大数据项目的部署实施都会依照这样的流程进行。事实上,即便是“小数据”驱动的数据分析项目也是按照这样的思路进行,只不过大数据时代的大数据特征让上述流程中的每一个环节(节点)都面临着巨大的挑战。

(1) 数据层的挑战

数据的复杂性是大数据项目开展伊始必须面对的难题。数据本身的复杂性包括数据类型的复杂性(数字、文本、图像、视频、音频等)、数据关系的复杂性(结构化数据、半结构化数据、非结构化数据等)、数据结构的复杂性(“属性—值”数据、关系型数据、流式数据等)、数据来源的复杂性(内部数据、外部数据、互联网数据等)、数据模式的复杂性(自媒体数据、富媒体数据、日志数据、传感器采集的数据等)等。

存在上述这些复杂性的同时,大数据的固有特点又让上述的复杂性更加复杂,比如数据的规模是巨大的(这意味着除了必须有海量的数据存取能力外,还需要有海量数据分析的能力)、数据的更新是快速的(这意味着除了必须有高效的数据存取能力外,还需要满足海量数据分析吞吐率的需求)、数据价值密度是稀疏的(这意味着除了需要有合理的数据存储模式外,在数据分析中还需要有稀疏数据处理的能力)、数据质量是混杂的(这意味着目标应用驱动的数据分析必须有能力甄别数据的真伪、优劣)等。

所有上述的复杂性直接导致数据分布的复杂^①、数据结构的复杂(网络数据的局部聚集特征、幂律特征、时序数据的动态演化和涌现等)、数据表示的尺度问题(指数据多尺度并存带来的表示尺度选择以及引发的度量问题)等。

(2) 存储层的挑战

上述数据的复杂性特点给数据的存取带来了极大的挑战,数据组织与管理的好坏是大数据项目开展成功与否的关键。大数据场景下的存储挑战在于:如何存储如此海量、复杂的数据,同时为后续的应用提供高效的存取、检索和更新。在数据存储层,除了要应付数据量大及数据种类多带来的挑战外,更棘手的挑战在于如何保证高效读取访问的同时具有高效更新能力。以主流的关系数据库为例,关系数据库的物理存储模式缺省值是堆文件(也就是随机文件),其目标是保证更新的效率和并发性。其面临的最大困难是既要保证查询快还需要保证更新快,而所有提高查询效率的技术(例如索引)都会降低更新效率,所有提高更新效率的技术都会降低查询效率。相比较而言,NoSQL数据库大都采用顺序存储,顺序文件能保证更

① 这里的分布不是指数据的存放位置,而是在数据统计分析意义下,数据彼此不是独立的而是关联的,样本的分布不代表数据的分布等。

高效的查询和基本的顺序插入。但是 NoSQL 的分布式特点使得基于不同 CAP 理念（一致性 Consistency、可用性 Availability、分区容忍性 Partition tolerance，简称 CAP，是分布式数据系统的三要素）的产品只能在不同的两个维度上寻得最优，而无法使得三者同时提高（关于 CAP 在 6.2.2 节中有详细介绍）。

（3）分析层的挑战

机器学习与数据挖掘（便于后续描述方便，可以先简单地理解“数据挖掘 = 机器学习 + 数据库”）是挖掘数据背后隐藏的知识和洞见的重要手段，而机器学习的本质是通过既有的数据建立一个“数据→目标”的映射函数 f' 。学习的目标是希望此 f' 能够与客观存在的那个函数 f 一致。因此，通常一个机器学习问题都牵涉两个重要的环节：如何通过既有的数据训练得到 f' ，以及如何评估 f' 与 f 的相似性。大数据场景下的学习问题是一个颇为棘手的复杂问题，其中的挑战包括如何学习以及如何判定学习建模的优劣。

1) 机器学习的一般流程是“特征提取→特征选择→学习建模→评估”，大数据场景下的数据的复杂性以及目标应用的多变性（传统的数据挖掘目标往往是确定的和单一的）使得面向目标应用的特征表示、特征提取和特征选择障碍很大。

2) 传统的机器学习评估是基于“测试集的样本分布和训练集的样本分布一致”的假设，一般的做法是在训练集上通过多倍交叉验证的方法（拿出一部分样本作为测试集，另一部分作为训练集）进行评估的。而在大数据时代这个假设无法成立，这就意味着在历史数据集上进行类似的测试评估，即便有很好的模型，其在新增的数据集上未必具有较好的表现，而如果拿新增的数据集作为测试集测试评估在历史数据集上的训练模型，显然也不公平。

3) 传统的机器学习一般是在封闭的数据集上，围绕某一个目标进行，而在大数据场景下，数据是开放的，且价值目标处于不断迭代中。显然传统的“数据封闭性假设”在大数据场景下无法成立，这意味着，必须寻求更多的方法和策略以响应大数据场景下数据层的开放性以及应用层的易变性特点。

一般而言，传统学习算法的算法复杂度（时间复杂度或空间复杂度）都会很大，在大数据场景下，这个问题会更加突出，如何应付这样的问题也是大数据学习中必须面对的挑战。或许还要考虑一个更为科学理论的问题：大数据场景下的所有问题一定都是可计算的吗？

（4）运维层的挑战

显然，大数据项目的建设是一个复杂的工程问题，需要各方资源的统一运筹。作为与目标用户交互的平台，大数据平台（系统）如何设计、实现、部署、实施是大数据落地应用中必须要考虑的问题。

一般而言，大数据项目所需要响应的目标需求是垂直化、细分化、扁平化和个性化的，如何高效地为用户提供目标指向的服务（计算服务和数据服务）不可避免牵涉到计算的负载均衡问题、个性化配置问题，这些是软件（系统）部署实施模式和体系架构问题，而如何调度各个计算资源为目标用户提供自适应的服务组合则显然是技术问题。一般而言，大数据项目涉及内部多个业务系统（模块）的集成和协作以及与外部业务系统（平台）的接口和联

动, 如何保证不同系统之间的有效集成和高效协同, 则涉及系统开发过程中的软件方法学、软件架构等。更为重要的是, 大数据软件(平台)的部署还只是大数据项目的开始, 在运维过程中, 如何有效地调度、配比各方资源进行平台的营运和运维则是大数据项目必须面对、但往往被忽略的问题。

(5) 架构层的挑战

如前所述, 大数据分析的核心环节是数据分析, 即机器学习与数据挖掘。鉴于现有机器学习与数据挖掘算法在算法复杂度上难以应对大数据分析中的挑战, 除了可以从理论上、技术上或者策略上去做革新或改良而外, 选择在合适的计算架构上进行算法的运行也是必须和必然的。基于分而治之的思想, 分布式计算是将大任务转化为若干可并行的小任务, 数据分析方面的挑战在于如何将现有的若干方法并行化, 两个简单的基本思路是将算法并行化或者在计算架构体系内实现算法的并行, 显然, 前者涉及算法模型的改造, 后者涉及分布式计算架构的改进。

分布式计算的一个典型优势是通过若干个计算性能不是那么优秀的分布式节点协作完成大任务的求解计算。此处一个隐含的改进思路是: 能否将每个计算节点的性能提高? GPU 计算是目前针对这一思路的一种解决方案, GPU 计算通过利用显卡闲置的计算资源大幅度提升每个节点的计算能力。在很多场合, 或许单个节点计算能力的提升已经能够满足实际需求, 即使不能满足, 将每个计算能力得到大幅提升的节点再纳入到分布式计算架构中, 其整体的计算性能一定是大为提高的。不过其中隐含的几个问题是: 哪些问题适合 GPU 计算? 如何将这种追求单个节点高性能计算的思路与不追求单个节点计算能力的分布式计算进行有效的集成从而达到严格意义上的计算性能最优?

4.2.2 隐式困难

除了上述的这些显式的挑战和困难外, 还有许多隐式的难题, 而这些难题也都是大数据项目部署、实施过程中必须面对和加以解决的。

(1) 数据在哪里? 凭什么能够拿到数据?

从大数据项目建设的主体来看, 数据的来源无非内部数据和外部数据两种, 内部数据指的是自营平台产生和采集的数据, 包括遗留系统的数据(一般指已经不再运维的被替代的系统数据以及纸质等媒体方式存储的不能被直接导入的数据); 外部数据指的是不在本单位内部, 而散布在其他单位服务器上或者互联网中的数据。从技术层面来看:

1) 本单位自营平台产生的数据或者遗留数据可以通过 ETL 来完成数据的整合, 若遗留数据是非数字化手段保存的, 则需要进行数字化转换以后导入。

2) 对于其他单位运营系统的数据, 可以通过商务合作的模式进行数据交换, 技术本质归根到底还是一种 ETL。

3) 对于互联网数据, 可以通过网络爬虫爬取数据或者通过某种商务合作, 通过专门的端口和 API 完成数据的获取。

尽管技术路线相对清晰,但在具体的大数据项目开展过程中,非技术因素往往成为数据获取的瓶颈。经常遇到的问题是:一般是通过购买或者利益交换获得其他单位(利益集团)的系统(平台)的数据(往往同一单位不同部门的数据集成也会出现类似的状况),那么,这笔数据值多少钱?这笔数据能给大数据项目开展带来多大的价值?大数据项目开展在数据获取层有多少预算?第一个问题牵涉到数据的估值问题,这个问题显然不是一个简单的技术问题;第二个问题则是要建立在对大数据项目开展的可行性研究、目标规划,对运维前景进行详细的策划、分析和运营等基础之上,这本身就是一个复杂的工程问题;第三个问题也是涉及工程管理的话题,更常见的场景是经费预算有限但需要购买大量外部数据源的数据,这给数据的获取带来的障碍往往会影响整个项目的开展进程。通过一个巧妙的让各方都有收益的商业模式设计,使得应该在前期数据采集的硬成本被分摊到未来的收益过程中,这成为一个可行的方案,但商业模式的设计,显然也不仅仅是技术,至少不仅仅是IT技术问题。

(2) 存储在哪里?有物理条件的制约吗?

数据的存储、管理和维护是大数据项目开展的重要环节。为了应对大数据量、高并发量、多数据类型等的挑战,对存储进行扩容几乎是一个不可避免的动作,而存储的扩容势必会提高整个大数据项目的经费预算,那么,摆在眼前的问题是:项目承担单位是否有足够的财务预算支撑持续不断的扩容需求?是否有相应配套的数据管理团队?数据管理团队的能力是否足够?目前流行的云存储是有效改善数据存储弹性扩容的模式。提供云存储的专业化团队在数据安全、备份及访问上的专业性往往会比本单位自建存储中心要优越得多,但问题是,作为大数据项目开展的主持单位,规则允许其将存储云化吗?如果采用云化存储,业务应用是否需要重新梳理?商业模式是否需要重新设计?

(3) 如何部署?应用目标环境支持吗?

任何一个大数据项目的持续推进和发展一定是因为以最恰当的方式完成了“数据→价值”的转换,这隐含着两个问题:目标应用有价值(有市场、直击用户“痛点”)、知识洞见有价值(能够从数据中发现知识,并且能够将知识应用在目标场景)。前者在于项目开展之初的目标定位及可行性研究(当然项目开展过程中也会出现迭代和反复),后者需要有技术支撑(从数据中发现知识)和应用能力(将知识应用在具体目标场景中)。排除技术因素不谈,大数据项目的持续运行必须要考虑到此应用场景是否有足够的数据支撑或者数据驱动的思维逻辑是否能够适应目标场景的需求。潜台词是:并不是所有的应用场景都适合应用大数据思维去构建系统平台。即便是大数据应用可行的应用场景,也要考虑到目标应用环境下,是否有匹配大数据项目建设需求的基础条件。

(4) 如何运维?有完备的运维团队吗?

前面提及,大数据项目系统(平台)部署后,需要有专门的运维团队对系统进行持续运维,这个运维团队需要有各方资源的共同介入,比如数据维护团队(负责数据源优化、数据优化、数据质量保障等)、法务支持团队(负责系统运维过程中牵涉的各类法律问题的协调和响应等)、商务支持团队(负责与各边的商务合作等)、技术支持团队(负责快速响应目标需

求的迭代更新及数据建模的持续跟进等)、系统维护团队(负责系统稳定运行保障等)等,更重要的是,如何有效地维系各边资源的协作也是一个重要的问题。因此,运维团队的建设及管理是大数据项目开展之初就必须要考虑并规划的内容。

4.2.3 评估思路

正因为大数据项目开展过程中存在上述的挑战和特点,对于大数据项目建设的成熟度(优或劣)可以从以下几个角度评估:

(1) 数据获取能力

任何一个大数据项目的开展,其重中之重是数据。能否以及如何将这些散布在不同数据源的数据高效(完备)获取并整合在同一数据平台上是评估大数据项目建设成熟与否的重要指标。需要特别注意的是,数据获取能力不仅在于从指定数据源中获得数据的能力,还在于有多大能力遴选合适的且足够的数据源,前者与技术和商务有关,后者则需要有充分的职业敏感度。另外一个需要注意的问题是:由于不同数据源的数据质量存在差异,如何在对不同数据源的数据进行数据剖析的基础上进行数据预处理,往往直接关系到数据的价值。

(2) 平台构建能力

任何一个大数据项目的构建都是基于基础设施层之上的大数据平台(事实上构建在基础设施层之上的平台层所承担的内容非常丰富,这里仅考虑与数据存取分析相关的大数据平台),然后基于此平台面向不同的目标应用开发相应的应用系统(平台)。因此,大数据平台的构建在整个大数据项目的开展过程中意义重大,一方面提供后续应用层高效透明访问数据的能力,另一方面需要提供后续应用层的分析计算能力。上述两种能力往往是以服务的形式加以封装并以服务的形式向后续应用层提供服务,简称数据服务和计算服务。应当注意到,所有的数据服务和计算服务都是针对既有数据进行的,这意味着大数据平台拥有多大的数据存储能力也是一个考量大数据平台的重要指标。

(3) 数据应用能力

任何一个大数据项目的最终价值体现在构建于大数据平台之上的应用(运维)平台在多大程度上匹配相应业务需求,这是数据转换为价值的重要标志。显然,应用平台的构建需要与业务目标紧密耦合,一方面需要对业务目标有准确的定位,另一方面需要对业务场景进行精准的建模。前者需要对应用极度敏感,后者需要有匹配的数据建模能力。另外需要注意的是,所有数据分析的结果(甚至中间结果)最终都要以合适的方式呈现给用户并与用户进行交互,同时细分需求的响应需要对数据进行不同尺度和维度的剖析,前者涉及数据呈现能力,后者涉及数据管理能力,这两者同样是数据应用能力评估的重要指标。

4.3 基础支撑技术

大数据项目的宗旨是“数据→价值”,这包含两个环节:一是“数据→知识”、二是“知

识→价值”，前者涉及数据采集、数据存储、数据建模（分析）等，后者涉及业务应用（运维）系统的设计与开发，本节重点讨论前者，重点说明每一个技术环节能够解决的问题以及针对相应的问题进行技术选型的应用提示。

4.3.1 数据采集

数据采集（Data Collection）是大数据项目开展的第一个环节，其主要职能是：从潜在数据源中获取数据并进行面向后续数据存储与管理以及数据分析与建模的预处理。上述的描述反映了数据采集环节中需要关注的几个重点。

（1）潜在数据源有哪些？

一般来说，大数据的来源可以分为三种（从大数据项目建设单位的角度而言）。

1) 平台自营型数据：大数据项目建设单位自主运维的软件平台产生的内部数据，这些数据除了包括软件平台生成的结构化或非结构化的数据以外，也包括在自主运维的传感器终端通过通信获取的数据。需要特别注意到的是，平台自营数据除了包括既有的自营平台外，也包括为当前大数据项目建设而有意开发（营运）的（新的、专门的）数据采集系统。

2) 其他主体运营数据：非大数据项目建设单位自主运维的软件平台产生的，而是存储在其他单位服务器的外部数据（此处的“外”是相对于大数据项目建设单位而言的）。某种意义上而言，这类数据的类型和格式与上述的平台自营性数据类似，只是这类数据往往要建立在某种商业模式意义下的交换而获得。

3) 互联网数据：散布于互联网中的数据，比如门户网站、社交平台、社区论坛等，从数据存储的物理位置来看，这类数据也是一种典型的其他主体运营数据。对于这类数据的获取可以通过某种商业模式意义下的交换获得，也可以通过网络爬虫自动爬取。

（2）如何从不同的数据源中获取数据？

针对平台自营型数据，数据与采集数据的工具都来源于平台内部。很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如 Hadoop 的 Chukwa、Cloudera 的 Flume、Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足每秒数百兆的日志数据采集和传输需求。有些平台自营型数据采集还与采集终端相关，这就涉及采集终端设计、通信协议、数据交换策略等。另外，与既有自营平台进行 ETL 也是一种常见的数据获取方法。

如 2.4 节中提及，ETL 是英文“Extract-Transform-Load”的缩写，较常用在数据仓库，但其对象并不限于数据仓库，是负责将分散的、异构数据源中的数据（如关系数据、平面数据文件等）抽取到临时中间层后，进行清洗、转换、集成，最后加载到数据仓库或数据集市，为联机分析处理、数据挖掘提供决策支持的数据。事实上，在大数据概念没有流行的时候，在数据交换场景下，ETL 应用已经相当普及。在大数据场景下，进一步凸显了 ETL 的意义。也正如前文 2.4 节中提及的，在大数据场景下，有学者提出应该将 ETL 改为 ELT，以响应大数据应用在数据采集层的挑战。

对于其他主体运营数据,在商务合作的基础上可以通过 ETL 实现数据的交换或者通过对方预留数据的访问接口获取数据;对于互联网数据,可通过网络爬虫实现数据的自动获取。

网络爬虫,又被称为网页蜘蛛,是一种网络数据采集方法,它按照一定规则自动抓取互联网数据的程序或者脚本。按照爬虫系统的软件架构,可以分为集中式爬虫系统和分布式爬虫系统,分布式爬虫系统是运行于集群之上的,集群中每一个节点都是一个集中式爬虫,分布式爬虫系统的体系结构有很多种,工作方式和存储方式也很多。但是,典型的分布式爬虫系统都采取主从方式的体系结构,即有一个主节点控制所有从节点执行抓取任务,这个主节点负责分配 URL,保证集群中所有节点的负载均衡。

(3) 如何预处理以适应后续的数据管理和分析?

将数据从不同的数据源采集到一起以后,需要对这些数据源进行必要的预处理,最终使得后续的数据分析得以有效进行。大致而言,数据预处理包括以下几个主要操作(不限于):

1) 清洗过滤。将数据中的噪声以某种技术或者既定策略去除并弥补缺失的数据。比如在互联网数据采集,一个 URL 指定的网页中,只有正文才是数据采集者感兴趣的。这意味着,我们要有相应的技术或者策略将网页中感兴趣的区域数据提取出来,而将其他反映网站模板结构的、广告信息的数据全部去除,借此降低后续存储负担的同时,也使得数据的质量得以提高,从而便于后续的分析。

2) 去重。将不同数据源的数据中的重复内容过滤,这种操作往往在互联网数据采集中尤其必要,比如针对新闻事件的分析,往往相同的新闻事件会在不同的网站上大量转载(有些转载时甚至是完全拷贝),在这种情况下,如果不把重复的数据过滤掉,一方面不利于后续的存储(存储负担过重),另一方面,重复的数据并没有更多的留存价值。一般而言,过滤类似重复内容时会以时间戳的方式记录重复的次数以及重复转载的 URL 明细(便于后续对这些 URL 进行评估和标签化)。

3) 建立数据的连接。从不同数据源获取数据的一个直接原因是希望通过互补的数据使得对目标对象的描述更加立体和具体,从而实现多数据源交叉复用的价值,这个动机的潜台词是:①在数据整合过程中,某数据源被采纳的一个必要条件是此数据源数据是对需要分析和研判的目标对象的直接描述,且对现有的描述具有互补性,这是在数据源的遴选时必须遵循的参考指标;②不同数据源的数据必须能够以某个对象为中心(节点)进行有效的关联和集成,彼此独立的数据没有集成的价值,这是在数据预处理中必须考虑的一个指标,牵涉到技术层和策略层的共同发力。

4) 特征化提取。有别于后续数据建模步骤必须对样本数据进行的特征表示、提取与选择,在数据预处理阶段进行的特征化提取一般专注于从原始数据(进行前置预处理后)中提取有语义的统计特征或者结构化特征,然后将这些特征作为该数据的一个标签存储(或者单独建立一个表,或者直接在原始数据的存储中增加一个字段)供后续的分析使用,比如从一段非结构化的法院公告文本中提取出有语义价值的原告、被告、判决时间等。

5) 标签化操作。标签化是大数据分析的一个典型策略和做法(后文有详细介绍),预处理环节中的标签化除了需要专注于将上述的特征化提取步骤获得的统计特性或者结构化语义信息提取(或者建模预测)出来作为数据的标签外,还需要考虑对各类数据源的置信度进行评估(即对数据源进行标签化),这样,当来自不同数据源的数据有冲突和歧义时,才能更好地进行综合研判。

4.3.2 数据存储

数据存储关注数据存在哪里以及如何存取,隐含着三个问题:

(1) 数据存在哪里?

毋庸置疑,物理上,数据一定是存在本地或者异地的磁盘上,不过需要引起注意的两个问题是:①物理存储模式是什么?②数据存储架构是什么?这两个问题往往是耦合的。一般的物理存储模式包括堆文件(随机文件)和顺序文件形式,前者能够保证更新的效率和并发性,后者能够保证后续更高效的查询和基本的顺序插入。数据的存储一般分为集中式和分布式,相比较于集中式存储,分布式存储在数据并发、负载均衡、数据安全等方面具有天然的优势,因此在大数据时代,分布式存储往往是天然的技术选型趋势。

(2) 数据如何存?

一般来说,数据的逻辑组织形式包括文件、数据库(包括关系型数据库、NoSQL),本质上无法说哪一种逻辑组织形式更好,只能说哪一种组织形式更适应于数据特征。经过几十年的发展,特别是数据库航母公司的强势推进,关系数据库成为事实上的业界标准。传统的关系数据库比较适合结构化数据的存储,追求的是“one size fits all”的目标,希望将用户从繁杂的数据管理中解脱出来。而在大数据时代,不同的应用领域在数据类型、数据处理方式以及数据处理时间的要求上有极大的差异,在实际的处理中几乎不可能有一种统一的数据存储方式能够应对所有场景。

随着技术的发展,适合大数据环境的新型数据库得到广泛的关注,其中,最典型的当属非关系型数据库 NoSQL。NoSQL 数据库不使用 SQL,同时 NoSQL 抛弃了关系模型并能够在集群中运行、不用事先修改结构定义也可自由添加字段,这些特征决定了 NoSQL 技术非常适用于大数据环境,从而成为业界翘楚,得到了迅猛的发展和推进。

(3) 数据如何取?

无论数据存在哪里以及如何组织,其最终目标都是为了后续的应用,最核心的问题是数据的存取。这牵涉两个问题:①如何高效快速地读取数据,即查询快;②如何高效快速地存储数据,即更新快。这两个目标都是用户需要的,但是往往存在冲突。以关系数据库为例,关系数据库的物理存储模式缺省值是堆文件,其目标是保证更新的效率和并发性。其面临的最大的困难是:所有提高查询效率的技术(例如索引)都会降低更新效率;而所有提高更新效率的技术,都会降低查询效率。近些年迅猛发展的 NoSQL 键值数据库大都采用顺序存储,其目标是能保证更高效的查询和基本的顺序插入,但更新困难。因此为了保障数据存取的高

效,“实时+批处理”往往是常用的一种策略。

在集中式数据环境中,为了解决计算能力,“实时+批处理”是常用的策略,即部分业务实时处理,部分业务批处理。例如银行的实时业务、行内和跨行清算批处理业务等。事实上,在分布式数据环境中,也存在类似的“实时+批处理”策略,比如 MapReduce 中 Map 可以理解成“实时”的映射,Reduce 可以理解成“批处理”的映射。

4.3.3 数据建模

大数据项目开展的核心是大数据分析,只有通过对数据的有效分析和建模,才能发掘数据背后的知识和洞见,而这种知识和洞见只有用于实际的目标应用场景,才能真正实现“数据→价值”的飞跃。普适的数据分析流程主要包括数据预处理、特征提取与选择、数据建模三个部分。

1) 数据预处理:现实世界中的数据易受噪声、缺失值和不一致性的侵扰,低质量的数据极有可能导致低质量的数据分析结果。数据预处理是对数据的第一步处理,主要包括数据清理、数据集成、数据规约、数据变换四种方法。数据清理可以用来清除数据中的噪声,纠正不一致性。数据集成将数据由多个数据源合并成一个一致的数据存储,如数据仓库。数据规约可以通过如聚集、删除冗余特征或聚类来降低数据的规模。数据变换可以把数据压缩到较小的区间,如 0 到 1,这可以提高挖掘算法的准确率和效率。

2) 特征提取与选择:数据建模的本质是建立“数据→目标”的模型,记为映射函数:

$$y = f'(x)$$

其中, x 是观察采样到的数据, y 是与输入对应的目标,函数 f' 就是拟建立的模型(或者认为是隐藏在数据背后的知识和洞见)。显然,作为输入的 x ,其承载着的是所观察采样到的原始数据,或许是文本,或许是图像,或许是视频等,无论是哪一种类型的数据,其计算量都会很大,维数灾难挑战突出。

维数灾难是指这样的现象:随着数据维度的增加,许多数据分析变得非常困难,数据在它所占的空间中越来越稀疏,结果是,对于高维数据许多分类和聚类算法的准确率都会急剧下降。去除无关维度往往会使接下来的数据分析或数据挖掘算法效果更好,可能的原因是:
①通过特征选择或者降维可以删除不相关的特征并降低噪声;
②通过特征选择和降维可以一定范围内消除维数灾难的影响。

因此,如果能从原始数据 x (高维空间) 提取出(映射成)一些线性无关的变量(低维空间)组成一个新的向量 x' ,在进行数据建模的时候,将原先建立“ $x \rightarrow y$ ”的映射关系改为建立“ $x' \rightarrow y$ ”的映射关系,则可以大范围降低计算量。从 x 提取 x' 的过程就是特征提取。特征提取的手段和方法有很多,有的从纯粹的数学角度做高维向量向低维向量的映射;有的从语义出发,有意识地提取具有高级语义的特征向量等。

尽管相比较于原始数据 x , x' 已经做了很大幅度的降维,但是在很多情况下(特别是用多

种特征提取方法提取了多组特征融合集成在一起的时候), 这个 x' 的维度还是很大, 因此一个自然的想法是, 能否把 x' 中对建模最优贡献的部分提取出来 (其他的部分抛弃, 以降低计算量), 这个过程就是特征选择。

3) 数据建模: 数据建模是从大数据中找出知识的过程, 常用的手段是机器学习和数据挖掘。如前文所述, 所谓数据挖掘可以简单地理解为“数据挖掘 = 机器学习 + 数据库”。从商业层次来说, 数据挖掘是企业按既定业务目标, 对大量企业数据进行探索和分析, 揭示隐藏的、未知的或验证已知的规律性, 并进一步将其模型化。从技术层次来说, 数据挖掘是通过分析, 从大量数据中寻找其规律的技术, 下面简单介绍几种常用的数据挖掘方法。

(1) 关联规则挖掘

顾名思义, 关联规则挖掘就是从数据背后发现事物 (务) 之间可能存在的关联或者联系。比如数据挖掘领域著名的“啤酒 - 尿不湿”的故事 (这个故事的真假不论) 就是典型的关联规则挖掘发现的有趣现象。在关联规则挖掘场景下, 一般用支持度和置信度两个阈值来度量关联规则的相关性 (关联规则就是支持度和信任度分别满足用户给定阈值的规则)。

所谓支持度 (Support), 指的是同时包含 X 、 Y 的百分比, 即 $P(X, Y)$; 所谓置信度 (Confidence) 指的是包含 X (条件) 的事务中同时又包含 Y (结果) 的百分比, 即条件概率 $P(Y|X)$, 置信度表示了这条规则有多大程度上可信。

关联规则挖掘的一般步骤是: 首先进行频繁项集挖掘, 即从数据中找出所有的高频项目组 (Frequent Item Set, 满足最小支持度或置信度的集合, 一般找满足最小支持度的集合); 然后进行关联规则挖掘, 即从这些高频项目组中产生关联规则 (Association Rule, 既满足最小支持度又满足最小置信度的规则)。

引用一个经典用例解释上述的若干概念, 使用的数据集如表 4-1 所示, 该数据集可以认为是超市的购物小票, 第一列表示购物流水 ID, 第二列表示每个流水 ID 同时购买的物品。

计算示例 1: 计算“如果 Orange 则 Coke 的置信度”, 即 $P(\text{Coke} | \text{Orange})$, 从上述的购物流水数据中可以发现, 含有 Orange 的交易有 4 个 (分别是 T1、T2、T3、T4), 在这 4 个项目中仅有两条交易含有 Coke (T1、T4), 因此,

$$P(\text{Coke} | \text{Orange}) = \frac{2}{4} = 0.5$$

表 4-1 超市购物流水

流水 ID	物品清单
T1	orange juice, coke
T2	milk, orange juice, window cleaner
T3	orange juice, detergent
T4	orange juice, detergent, coke
T5	window cleaner

计算示例 2: 计算在所有的流水交易中“既有 Orange 又有 Coke 的支持度”, 即 $P(\text{Orange}, \text{Coke})$, 从上述的购物流水数据中可以发现, 总计有 5 条交易记录 (T1、T2、T3、T4、T5), 既有 Orange 又有 Coke 的记录有两条 (T1、T4), 因此,

$$P(\text{Orange}, \text{Coke}) = \frac{2}{5} = 0.4$$

上述两个计算示例总结出的关联规则是：如果一个顾客购买了 Orange，则有 50% 的可能购买 Coke。而这样的情况（即买了 Orange 会再买 Coke）会有 40% 的可能发生。

（2）分类

在数据挖掘领域，分类可以看成是从一个数据集到一组预先定义、非交叠类别的映射过程，其中映射关系的生成以及映射关系的应用就是数据挖掘分类方法主要的研究内容。这里的映射关系就是我们常说的分类函数或分类模型（分类器），映射关系的应用就对应于我们使用分类器将数据集中的数据项划分到给定类别中的某一个类别的过程。

我们人是怎么区分一个人是男性还是女性的问题就是一个典型的分类问题。在我们的脑中早就建立了男人和女人的模型，每当我们遇到一个陌生人的时候，我们的大脑就获取到了这个人的特征信息，通过大脑中的模型去将这个人归类到男性或者女性的类别中（当然人的大脑神经系统处理这个问题时的流程往往比我们这里叙述的复杂得多）。但是我们大脑中的模型是怎么建立的呢？是生来就有的吗？很明显不是。我们大脑建立模型的过程都是从过去的经验中总结积累出来的，并在实践过程中不断地修正或扩充（从数据中学习）。

分类是从历史的特征数据中推导出特定对象的描述模型，用来对未来数据进行预测和分析。分类算法是解决分类问题的方法，是数据挖掘、机器学习和模式识别中一个重要的研究方向。分类算法通过对已知类别训练集的分析，发现分类规则，以此预测新数据的类别。分类算法的应用非常广泛，如银行风险评估、客户类别分类、文本检索和搜索引擎分类、安全领域中的入侵检测以及软件项目中的应用等。常见的分类算法有决策树、贝叶斯分类、人工神经网络、K-近邻等。

（3）聚类

聚类是将数据聚集到不同的类或者簇的过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。

机器学习的分类体系有多种，从学习的方式分为示例学习、类比学习、分析学习等，其中示例学习是关注度和研究热度最高的一个方向。示例学习又有很多分类方法：从学习的主动性方面分类，可以分为主动学习和被动学习；从训练过程启动的时间分类，可以分为急切式学习和懒惰式学习等。最常见的对示例学习的分类是监督学习、非监督学习和强化学习，这是从训练样本的歧义性来进行分类的。

监督学习通过对具有概念标记的训练集进行学习，以尽可能正确地对训练集之外的示例的概念标记进行预测。这里所有训练示例的概念标记都是已知的，因此训练样本的歧义性最低。非监督学习通过对没有概念标记的训练示例进行学习，以发现训练示例中隐藏的结构性知识。这里的训练示例的概念标记是不知道的，因此训练样本的歧义性最高。强化学习通过

对没有概念标记、但与一个延迟奖赏或效用（可视为延迟的概念标记）相关联的训练示例进行学习，以获得某种从状态到行动的映射。这里本来没有概念标记的概念，但延迟奖赏可被视为一种延迟概念标记，因此其训练样本的歧义性介于监督学习和非监督学习之间。

半监督学习是用未标记的样本来辅助对已标记样本的学习，半监督学习中的歧义性是人为的（与监督学习、非监督学习、强化学习等天生的歧义性完全不同），这在解决工程问题上往往是需要的、有用的，毕竟对大量样本进行标记的代价可能是极为昂贵的。

与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式学习。聚类分析是一种探索性的分析，在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用方法的的不同，常常会得到不同的结论。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。而且聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合做进一步的分析。同时，聚类分析还可以作为其他算法（如分类和定性归纳算法）的预处理步骤。常见的聚类算法有 K-Means、层次聚类、基于密度的聚类等。

事实上，面向不同的应用目标，数据挖掘方法还有很多，此处不再一一赘述。

4.3.4 计算架构

大数据的复杂性给大数据分析带来的挑战至少有两点：

- 1) 如何响应数据类型的复杂性给数据的理解、建模带来的挑战。
- 2) 如何用更快的计算效率响应数据的海量、并行及快速更新的特性。

前者的挑战需要研发新型的理论、算法、技术，而后者需要所有的算法、技术（改进）必须依赖合适的高性能计算架构。

目前用于高性能计算的策略有两类。

1. 将复杂的任务“分而治之”

将复杂的任务“分而治之”，引入分布式计算架构以提升计算性能。分布式计算的基本出发点在于不单纯追求每个计算节点的计算性能有多强，而是通过更多的计算能力不是那么强的计算节点，通过某种合适的策略达到整体计算性能极大提升。根据不同的分布式策略和目标，目前已经有包括 Hadoop、Spark、Storm 等主流分布式计算产品，这些产品都是基于 MapReduce 思路的。

以下简单介绍有关 MapReduce 的相关内容。

MapReduce 是大数据时代炙手可热的一个技术名词，其设计初衷是通过大量廉价的服务器实现大数据的并行处理。它对数据一致性要求不高，其突出优势是具有扩展性和可用性，特别适用于海量的结构化、半结构化及非结构化数据的混合处理。它是 Google 在 2004 年提出的

一个软件架构, 包括两个基本的过程: Map 和 Reduce (Map 和 Reduce 的主要思想是从函数式编程语言及向量编程语言借鉴而来的)。

MapReduce 的基本过程如图 4-2 所示。

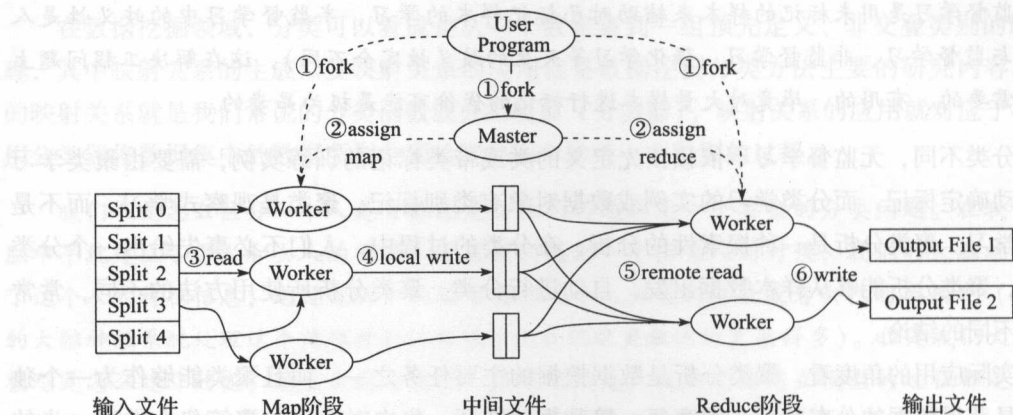


图 4-2 MapReduce 流程图

如图 4-2 所示, MapReduce 是从 User Program 开始的, User Program 链接了 MapReduce 库, 实现了最基本的 Map 函数和 Reduce 函数, 执行的顺序分别如下:

1) MapReduce 库先把 User Program 的输入文件划分为 M 份 (M 由用户定义), 每一份通常有 16MB 到 64MB, 如图 4-2 所示分成了 Split 0 ~ 4; 然后使用 fork 将用户进程拷贝到集群内其他机器上, 见图 4-2 步骤①。

2) User Program 的副本中有一个称为 Master, 其余称为 Worker (数量可由用户指定), Master 负责为空闲 Worker 分配作业 (Map 作业或者 Reduce 作业), 见图 4-2 步骤②。

3) 被分配了 Map 作业的 Worker, 开始读取对应分片 (Split) 的输入数据, Map 作业数量是由 M 决定的, 与 Split 一一对应; Map 作业从输入数据中抽取键值对, 每一个键值对都作为参数传递给 map 函数, map 函数产生的中间键值对被缓存在内存中, 见图 4-2 步骤③。

4) 缓存的中间键值对会被定期写入本地磁盘 (见图 4-2 步骤④), 而且被分为 R 个区 (R 由用户定义), 将来每个区会对应一个 Reduce 作业; 这些中间键值对的位置会被通报给 Master, Master 负责将信息转发给 Reduce worker。

5) Master 通知分配了 Reduce 作业的 Worker 它负责的分区在什么位置 (肯定不止一个地方, 每个 Map 作业产生的中间键值对都可能映射到所有 R 个不同分区), 当 Reduce worker 把所有它负责的中间键值对都读过来后 (见图 4-2 步骤⑤), 先对它们进行排序, 使得相同键的键值对聚集在一起。

6) Reduce worker 遍历排序后的中间键值对, 对于每个唯一的键, 都将键与关联的值传递给 reduce 函数, reduce 函数产生的输出会添加到这个分区的输出文件中, 见图 4-2 步骤⑥。

7) 当所有的 Map 和 Reduce 作业都完成了, Master 唤醒 User Program, MapReduce 函数调

用返回 User Program 代码。

所有执行完毕后, MapReduce 输出放在了 R 个分区的输出文件中 (分别对应一个 Reduce 作业)。用户通常并不需要合并这 R 个文件, 而是将其作为输入交给另一个 MapReduce 程序处理。整个过程中, 输入数据和输出数据在底层分布式文件系统 (GFS) 中进行, 而中间数据是放在本地文件系统的。需要注意 Map/Reduce 作业和 map/reduce 函数的区别: Map 作业处理一个输入数据的分片, 可能需要多次调用 map 函数来处理每个输入键值对; Reduce 作业处理一个分区的中键值对, 期间要对每个不同的键调用一次 reduce 函数, Reduce 作业最终也对应一个输出文件。

基于上述 MapReduce 的基本原理, 不同的公司根据不同的思路开发了不同的分布式计算架构的产品, 主要有: Apache Hadoop、Spark、Storm 等。

(1) ApacheHadoop

Apache Hadoop 是一款支持数据密集型分布式应用的开源软件框架, 其最核心的设计就是分布式文件系统 (Hadoop Distributed File System, HDFS) 和 MapReduce。HDFS 为海量的数据提供了存储, MapReduce 则为海量的数据提供了计算。

HDFS 是基于 Google 的 BigTable 实现的。BigTable 是一个为管理大规模结构化数据而设计的分布式存储系统, 可以扩展到 PB 级数据和上千台服务器。BigTable 看起来像一个数据库, 采用了很多数据库的实现策略, 但是 BigTable 并不支持完整的关系型数据模型, 而是为客户端提供了一种简单的数据模型, 客户端可以动态地控制数据的布局和格式, 并且利用底层数据存储的局部性特征。BigTable 将数据统统看成无意义的字节串, 客户端需要将结构化和非结构化数据串行化再存入 BigTable。

(2) Spark

Spark 是一个由加州大学伯克利分校 AMP 实验室 (Algorithms, Machines, and People Lab) 开发的基于内存计算的开源集群计算系统, Spark 拥有和 Hadoop 相似的开源集群计算环境, 不同之处在于:

首先, Spark 是为集群计算中的特定类型的工作负载而设计的, 即那些在并行操作之间重用数据集 (比如机器学习算法) 的工作负载。为了优化这些类型的工作负载, Spark 引进了内存集群计算的概念, 可在内存集群计算中将数据集缓存在内存中, 以缩短访问延迟。

其次, Spark 还引进了名为弹性分布式数据集 (RDD) 的抽象, RDD 是分布在一组节点中的只读对象集合。这些集合是弹性的, 如果数据集一部分丢失, 则可以依赖容错机制 (该机制允许基于数据衍生过程重建部分数据集的信息) 对它们进行重建。

最后, 尽管创建 Spark 是为了支持分布式数据集上的迭代作业, 但与 Hadoop 类似, Spark 支持单节点集群或多节点集群。对于多节点操作, Spark 通过名为 Mesos 的第三方集群管理器可以实现 Spark 在 Hadoop 文件系统中并行运行。

(3) Storm

Hadoop (包括在 Hadoop 基础上发展起来的 Spark) 在本质上是一个批处理系统, 数据被

引入到 HDFS 并分发到各个节点进行处理, 当处理完成时, 结果数据返回到 HDFS 供始发者使用。毋庸置疑, Hadoop 是大数据分析领域无可争辩的王者 (虽然不断出现的新计算模型不断对其发起挑战), 但其仅适用于大规模静态数据的批处理场合, 而如果数据来源于高度动态的实时信息 (流) 时, Hadoop 就不再适用。针对此局限, Nathan Marz 推出的 Storm 提供了一个解决方案。

Storm 是一个开源的实时计算系统, 它提供了一系列的基本元素用于计算, 包括 Topology、Stream、Spout、Bolt 等。

在 Storm 中, 一个实时应用的计算任务被打包作为 Topology 发布, 这同 Hadoop 的 MapReduce 任务相似。但是有一点不同的是: 在 Hadoop 中, MapReduce 任务最终会在执行完成后结束; 而在 Storm 中, Topology 任务一旦提交后永远不会结束, 除非用户显式停止任务。

计算任务 Topology 是由不同的 Spouts 和 Bolts 通过数据流 (Stream) 连接起来的图, 如图 4-3 所示。

如图 4-3 所示, 在一个 Topology 中包含两个基本元素。

1) Spout: Storm 中的消息源, 用于为 Topology 生产消息 (数据), 一般是从外部数据源不间断地读取数据并给 Topology 发送消息 (tuple 元组)。

2) Bolt: Storm 中的消息处理器, 用于为 Topology 进行消息的处理, Bolt 可以执行过滤、聚合、查询数据库等操作, 而且可以一级一级地进行处理。

最终, Topology 会被提交到 Storm 集群中运行, 也可以通过命令停止 Topology 的运行, 将 Topology 占用的计算资源归还给 Storm 集群。

数据流 (Stream) 是 Storm 中对数据进行的抽象, 它是时间上无界的 tuple 元组序列。在 Topology 中, Spout 是 Stream 的源头, 负责为 Topology 从特定数据源发射 Stream; Bolt 可以接收任意多个 Stream 作为输入, 然后进行数据的加工处理过程, 如果需要, Bolt 还可以发射出新的 Stream 给下级 Bolt 进行处理。图 4-4 显示了一个 Topology 内部 Spout 和 Bolt 之间的数据流关系:

Topology 中每一个计算组件 (Spout 和 Bolt) 都有一个并行执行度, 在创建 Topology 时可以进行指定, Storm 会在集群内分配对应并行度个数的线程来同时执行这一组件。同时, Storm 提供了若干种数据流分发 (Stream Grouping) 策略用来解决 Spout 和 Bolt 两个元组之间发送 tuple 元组的问题。在 Topology 定义时, 需要为每个 Bolt 指定接收什么样的 Stream 作为其输入 (注: Spout 并不需要接收 Stream, 只会发射 Stream)。目前 Storm 中提供了以下 7 种 Stream Grouping 策略: Shuffle Grouping、Fields Grouping、All Grouping、Global Grouping、Non Grouping、Direct Grouping、Local or Shuffle Grouping, 具体介绍此处不再赘述。

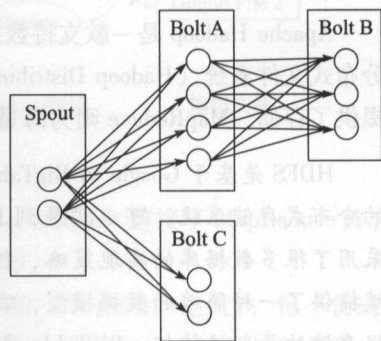


图 4-3 Topology 示意图

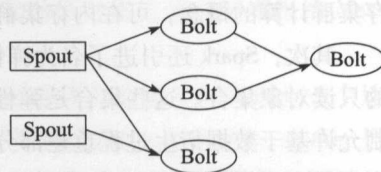


图 4-4 Topology 数据流示意图

通过以上介绍可以看出, Storm 中 Stream 的概念是 Topology 内唯一的, 只能在 Topology 内按照“发布-订阅”方式在不同的计算组件 (Spout 和 Bolt) 之间进行数据的流动, 而 Stream 在 Topology 之间是无法流动的, 这个特点需要在应用的过程中加以留意。

2. 充分提升和挖掘单个计算节点的计算性能

比如通过对计算主机进行 CPU、内存、硬盘等的扩容尝试增加单个计算节点的计算性能, 显然, 这已不是纯粹的技术层次的问题。值得关注的技术流是 GPU 技术, 通过充分利用单个计算节点的 (显卡) 剩余能力, 达到对单台计算机大幅提升计算性能的目的。

GPU 是 Graphics Process Unit 的简称, 也就是图形处理器, 就这个意义而言, GPU 是显卡的核心部件, 早期 GPU 仅是用于硬件加速 (部分功能从 CPU 分离), 仅能起到 3D 图像处理的加速作用, 不具有软件编程特性。随着技术的发展, GPU 的可编程性逐渐增强, 特别是在 2006 年 NVIDIA 与 ATI 分别推出了 CUDA (Computer Unified Device Architecture) 编程环境和 CTM (Close To the Metal) 编程环境, 使 GPU 通用计算编程的复杂性、可编程性、功能、性能不断提升和完善, GPU 已演化为一个新型可编程高性能并行计算资源。

CUDA 是显卡厂商 NVIDIA 推出的通用并行计算架构, 该架构使 GPU 能够解决复杂的计算问题。它包含了 CUDA 指令集架构以及 GPU 内部的并行计算引擎。开发人员使用 C 语言 (随着版本的升级, 逐步支持 C++、Fortran 等) 基于 CUDA 架构编写的程序可以在支持 CUDA 的处理器上以超高性能运行。对应用程序开发人员来说, 这是一个巨大的市场。

CPU 的设计目标是: 具备更强的适应不同数据类型的通用计算能力, 同时又要能够进行因逻辑判断而引入的大量分支跳转和中断处理, 而 GPU 面对的则是类型高度统一的、相互无依赖的大规模数据和不需要被打断的纯净计算环境。两者设计目标的不同导致了两者呈现出非常不同的架构, 见图 4-5 (该图摘自 NVIDIA CUDA 文档, 略加修改)。

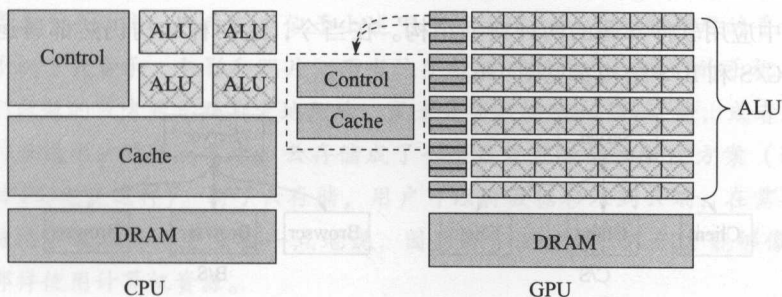


图 4-5 CPU 和 GPU 架构示意图

由图 4-5 可见, GPU 采用了数量众多的计算单元和超长的流水线, 但只有非常简单的控制逻辑并省去了 Cache。而 CPU 不仅被 Cache 占据了大量空间, 而且还有复杂的控制逻辑和诸多优化电路, 相比之下计算能力只是 CPU 很小的一部分。与 CPU 擅长逻辑控制和通用类型数据运算不同, GPU 擅长的是大规模并发计算, 尤其是计算密集型场合。

由于架构的原因，CPU 计算性能的提高一直都从提升制程工艺和主频上着手，但这显然是有壁垒的。如果说 CPU 是一条单车道公路，一次只能走一辆车的话，GPU 就是拥有多个车道的高速公路，正是大量并行的结构使得其在浮点运算方面拥有了非常快的速度，而在 GPU 基础上发展起来的 GPGPU（通用图形处理器）很好地继承了这一优点，其具有比 CPU 高一个数量级的浮点性能，在注重运算速度的高性能领域被格外看好。因此将并行运算能力更为强大的 GPU 释放出来，并与 CPU 相结合进行运算，开始成为超级计算机领域一个新的解决方案。

目前主流的 GPU 结构有两种，一个是基于流处理器阵列的主流 GPU 结构，以 NVIDIA 的 GeForce8800GTX 和 ATI 的 HD 2900 为代表；另外一个基于通用计算核心的 GPU 结构，以 Intel Larrabee 为代表。前者相对于后者具有更高的聚合计算性能，而后者则在可编程性上具有更大的优势。

4.4 高级支撑技术

大数据能够得到“政产学研商用”各界的普遍关注，一个重要的原因是因为人们一致认为数据背后隐含着价值。而如何从数据中挖掘出价值除了纯粹的技术因素外，还与具体的应用模式、商业模式等非技术因素相关，有关应用模式梳理和商业模式梳理的相关内容会在后面的章节中具体探讨，本节尝试在上节的基础上介绍与大数据项目部署实施直接相关的云计算话题。

4.4.1 云计算背景

C/S（Client/Server，客户机/服务器）和 B/S（Browser/Server，浏览器/服务器）是信息技术发展过程中应用软件部署的两种典型结构。在当今，这两种结构仍然部署运行在不同的应用场景中，C/S 和 B/S 结构参见图 4-6。

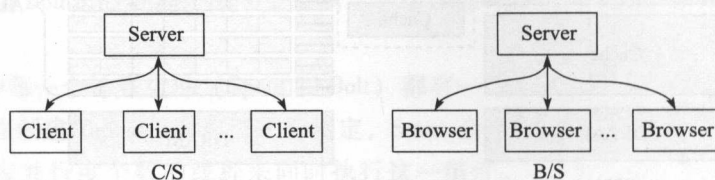


图 4-6 C/S 与 B/S 对比

从图 4-6 中好像看不出 C/S 和 B/S 的差别来，因为其结构都是建立在独立的客户终端和服务端两层分离基础上的。C/S 强调的是将计算负载均摊在客户机端和服务端，因此基于 C/S 结构的应用软件必须在服务器和客户机端均部署安装；而 B/S 强调的是瘦客户端模式，客户端不用安装专门的应用软件，而只需安装浏览器软件即可，所有的业务逻辑、数据存取均在服务器端完成。看起来的表象差别事实上引发了技术层次和应用层次的显著不同，比如：

基于 B/S 结构的系统可以做到一次安装（在服务器端），各个终端（只要装了浏览器软件）均可运行，相比较于基于 C/S 结构的系统，不仅要在服务器端安装系统，也要在终端安装系统，这意味着版本的升级对于运维工程师而言几乎是一个“灾难”。当然两者的诸多区别此处不做赘述，因为这不是本文的重点，本文关心的是这两种结构的一个共性特征而必然引发的“云计算”这一应用模式的诞生。

无论是 C/S 结构，还是 B/S 结构，服务器的选型和采购都是一个颇为棘手的问题，对于项目主（甲方）而言也是一个不可小觑的“黑洞”。比如项目筹建之初，项目主需要根据当时的业务逻辑和计算负载需求采购相应的服务器、存储等基础设施以及数据库、操作系统等软件授权，甚至还要（一般都必须）建立专门的机房并配备专门的技术人员，而这不是一成不变的，当企业规模变大引发项目规模变大，基础设施的扩容以及各种软件的升级（往往意味着继续为了授权而缴付费用）就成了不二选择，而所有的这一切都意味着甲方必须投入巨大（甚至不可预料的）的铺底资金。一个极端情况是：某些业务场景，可能仅仅是某一段时间内对基础设施的扩容有巨大的需求，而过了这段时间，扩容的基础设施其实就是一种浪费。以 12306 这样的订票网站为例，“春运”期间其服务器负载尤其巨大，而过了这段时间，往往负载要锐减很多。

在很多技术储备均得以迅猛发展的基础上，基于 SOC/SOA (Service-Oriented Computing/Service-Oriented Architecture) 框架的云计算 (Cloud Computing) 应用模式受到了越来越广泛的关注，并且其普及度也在逐步地深入，而这种应用模式很好地响应了上述的需求：厂商将硬件资源（服务器、存储、CPU、带宽等）和软件资源（应用软件、集成开发环境等）以服务的形式按需分配给用户，用户仅需支付服务费而无须如从前一样购买基础设施和应用软件授权等。

对于消费者个人而言，也存在类似于上述甲方的困境。以每人都会用的手机为例，每天我们都会用手机下载音乐、电影和照片，而当这些数据不断增长，我们的手机存储难以维系的时候，我们能做的或者是忍痛割爱地删除，或者是将数据移入 PC 存储，或者是增加更多的存储……针对普通用户的痛点需求，云存储成了一个匹配度很好的解决方案（这不仅是对移动终端，对 PC 也是这样）。有了云存储，用户可以将数据移入到云端，在需要的时候直接在云端进行访问和读取……更直接一点地说，因为有了云计算，人们才能够像使用水、电、煤气和电话那样使用计算机资源。

4.4.2 云计算定义

针对“云计算”，目前还没有统一的定义（不同的学者，尤其是不同的厂家，出于不同的利益诉求，总是进行一些彼此有别的定义），较为一致的观点是：云计算是以虚拟化技术为基础，以网络为载体提供基础架构、平台、软件等服务的形式，整合大规模可扩展的计算、存储、数据、应用等分布式计算资源进行协同运作的超级计算模式。上述的定义中涉及几个关

键要点, 分别如下:

(1) 云计算的服务形式

在云计算中的应用中, 云计算可划分为三个服务层, 即以下三个服务集合。

1) SaaS(Software as a Service): 软件即服务, SaaS 是一种通过 Internet 提供软件的模式, 用户无须购买软件, 而是向提供商租用基于 Web 的软件来管理企业经营活动, 比如 163 信箱、微博、微信等。

2) PaaS(Platform as a Service): 平台即服务, PaaS 实际上是指将软件研发的平台作为一种服务, 以 SaaS 的模式提交给用户。因此, PaaS 也是 SaaS 模式的一种应用。但是, PaaS 的出现可以加快 SaaS 的发展, 尤其是加快 SaaS 应用的开发速度。

3) IaaS(Infrastructure as a Service): 基础设施即服务。消费者通过 Internet 可以从完善的计算机基础设施获得服务, 如硬件服务器租用等。

(2) 云计算相关技术

云计算是网格计算(Grid Computing)、分布式计算(Distributed Computing)、并行计算(Parallel Computing)、效用计算(Utility Computing)、网络存储(Network Storage)、虚拟化(Virtualization)、负载均衡(Load Balance)等传统计算机和网络技术发展融合的产物。本质上, 云计算就是综合利用上述技术, 将大量用网络连接的计算资源统一管理和调度, 构成一个计算资源池向用户提供按需服务, 核心技术是虚拟化技术。

1) 网格计算: 分布式计算的一种, 由一群松散耦合的拥有计算能力的节点之间形成联盟, 共同解决大规模计算的问题, 是基础 IT 资源联合共享模式的运用。云计算和网格计算都能够提高 IT 资源的利用率。但是云计算侧重于 IT 资源的整合, 整合后按需提供 IT 资源; 网格计算侧重于不同组织间计算能力的连接。云计算依靠 IT 资源供给的灵活性, 革新了 IT 产业的商业模式, 是基础 IT 资源外包商业模式的典型运用。可以粗放地认为, 网格计算的目标是高性能计算, 而云计算的目标是按需计算。

2) 分布式计算: 是相对于集中计算而言的, 是指将一个需要非常巨大的计算能力才能解决的问题分成许多小的部分, 然后把这些部分分配给许多计算机进行处理, 最后把这些计算结果综合起来得到最终的结果, 其基本思路是“分而治之、合作求解”。

3) 并行计算: 是相对于串行计算而言的, 是指同时使用多种计算资源解决计算问题的过程, 可分为时间上的并行和空间上的并行。前者是指流水线技术, 后者是指用多个处理器并发地执行计算, 并行计算的主要目的是快速解决大型且复杂的计算问题。

4) 效用计算: 是一种提供服务的模型, 在这个模型里服务提供商产生客户需要的计算资源并提供基础设施管理, 根据具体应用对 IT 资源进行打包, 按照计算、存储分别计量费用, 而不是仅仅按照速率进行收费。

5) 网络存储: 是基于数据存储的一种通用网络术语, 基本存储结构包括直连式存储(Direct Attached Storage, DAS)、网络存储设备(Network Attached Storage, NAS)和存储网络(Storage Area Network, SAN)等, 详细参见本书第 6 章。

6) 虚拟化: 是一种资源管理技术, 将计算机的各种实体资源, 如服务器、网络、内存及存储等进行抽象、转换后呈现出来, 使用户以比原本组态更好的方式来应用这些资源, 包括将单个资源划分成多个虚拟资源的裂分模式, 也包括将多个资源整合成一个虚拟资源的聚合模式。虚拟化技术根据对象可分成存储虚拟化、计算虚拟化、网络虚拟化等, 计算虚拟化又分为系统级虚拟化、应用级虚拟化和桌面虚拟化。通过虚拟化实现资源“空分、时分”复用, 提高资源的利用率。因此虚拟化技术是云计算中最为核心的关键技术。

7) 负载均衡: 是建立在现有网络结构之上的一种廉价有效透明的方法, 用以扩展网络设备和服务器的带宽、增加吞吐量、加强网络数据处理能力、提高网络的灵活性和可用性。具体包括软件负载均衡(在一台或多台服务器相应的操作系统上安装一个或多个附加软件实现负载均衡)、硬件负载均衡(直接在服务器和外部网络间安装独立于操作系统的负载均衡设备, 通过在负载均衡设备上配置不同的负载均衡策略达到负载均衡需求)、本地负载均衡(对本地的服务器群实现负载均衡)、全局负载均衡(对分别放置在不同地理位置、不同网络结构的服务器群间实现负载均衡)等。

8) 自主计算: 通过现有的计算机技术来替代人类部分工作, 使计算机系统能够自调优、自配置、自保护、自修复, 以技术管理技术方式提高计算机系统的效率并降低管理成本。

需要说明的是, 云计算所涉及的相关技术远不止上述提到的八类, 且每一类技术实现的思路和手法也各有千秋, 此处不一一赘述。

4.4.3 云计算本质

云计算的本质是一种基于互联网的应用模式, 该应用模式有效地嫁接了计算资源提供者 and 计算资源消费者(需求者)的双边关系, 并各取所需, 它是一种技术和商业模式的双重创新。从技术层面来看, 云计算就是虚拟化技术和集群管理技术的合体, 目标是为计算资源消费者提供按需分配的计算资源。从商业模式上来看, 计算资源消费者(需求方)从早期的“购买计算资源”转向“购买计算资源服务”, 从而在获得相应的计算资源的同时, 降低计算资源方面的投入成本; 计算资源提供者通过对自有计算资源的虚拟化实现计算资源的“空分、时分”复用, 实现资源的有效利用, 从而获得更多收益。图4-7 简要显示了此种关系。

1) 图4-7a 是传统的消费者和计算资源的关系, 每个计算资源消费者所需要的计算资源必须要自行采购, 如前所述, 此种模式难以应付(尤其是难以应付临时的)基础设施扩容等方面的场景。

2) 图4-7b 是基于租赁模式的一种计算资源分配方式, 大概的思路是: 作为计算资源的消费者无须采购相应的资源, 而以租赁的方式向计算资源的拥有者租赁, 所支付的租赁费一般远小于采购成本, 而计算资源的提供方(租赁方)通过获取租赁费以及其他一些盈利模式最终受益(这种意义上的租赁方往往是大批量的采购, 会获得比每个消费者单独采购低得多的议价, 更何况还有其他一些盈利模式, 此处不赘述)。

3) 图4-7c 是在租赁模式基础上发展起来的云租赁模式, 该模式综合考虑了不同消费者对

计算资源的依赖具有典型的“空分、时分”特点，以虚拟化为基础，采用分布式计算和存储，结合优化的硬件，通过集群化运维管理系统，实现计算、存储、网络等资源的动态分配及部署，真正为每个消费者提供他们所需的（确切地说是按需分配的）计算资源（虚拟的）。显然在这种模式的指导下，关键技术是虚拟化，对双边的好处是：计算资源拥有方并不需要采购那么多的资源，并且有了一定的资源后，可以服务更多的消费者；而消费者并不需要支付庞大的费用就能够获得所需的计算资源，是一种互利双赢。

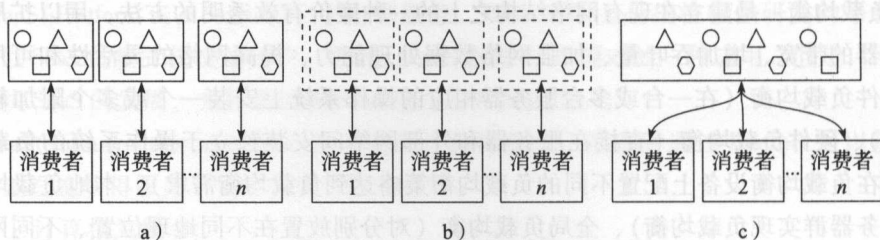


图 4-7 基于云计算的商业模式变迁

综上所述，云计算具有如下几个典型的特点：

1) 规模超大。一般而言，云计算服务商提供的“云”具有相当的规模，能够没有悬念地保证消费者的“时分、空分”共享，即便在极端情况下，所有消费者对计算资源的需求没有“时分、空分”的特征，也能够保证消费者对计算资源的需求（因为“云”的规模是可以动态伸缩的，借此满足应用和用户规模增长的需要）。

2) 费用低廉。“云”所具备的很多特性使其性价比极高，包括：“云”的特殊容错措施使得可以采用极其廉价的节点来构成云；“云”设施可以建在电力资源丰富的地区，从而大幅降低能源成本；“云”的自动化管理使数据中心管理成本大幅降低；“云”的公用性和通用性使资源的利用率大幅提升。而这种低成本也直接反馈在消费者使用和租赁“云”方面的成本开销大幅降低。

3) 按需分配。“云”是一个庞大的资源池，消费者可以根据自己的业务需求购买定制的虚拟资源服务，通过云快速地提供虚拟机或物理机器，迅速部署和增加工作负载，同时实时监控资源使用情况，在需要时重新平衡资源分配。

4) 通用性高。云计算作为一个虚拟化的计算机资源池，不针对特定的应用，在“云”的支撑下可以构造出千变万化的应用，同一片“云”可以同时支撑不同的应用运行，它可以托管多种不同的工作负载，包括成批的后端作业和面向用户的交互式应用程序。

5) 安全可靠。“云”使用了数据多副本容错、计算节点同构可互换等措施来保障服务的高可靠性，同时云计算支持冗余、自我恢复且提供具有高可扩展性的编程模型，使工作负载能够从多种不可避免的硬件/软件故障中进行恢复，从而保障使用云计算比使用本地计算（机）更加可靠。

6) 虚拟化。云计算支持用户在任何时间、任何地点、使用任何终端获取应用服务。由于

应用在“云”中，所以用户无须了解和顾虑这种计算具体在哪个地方运行。

根据云计算服务对象和服务性质，云计算的部署模式包括公有云、私有云和混合云三种，分别介绍如下：

(1) 公有云

公有云通常是指第三方提供商将搭建好的云资源池放到 Internet 上，所有有使用权限的用户都可以按需使用（免费或者收费），公有云部署模式被公认为是云计算的主要形态。

公有云在国内外的的发展如火如荼，根据市场参与者类型分类，国内的公有云可以分为四类：第一类为传统电信基础设施运营商，包括中国移动、中国联通和中国电信；第二类是政府主导下的地方云计算平台；第三类是互联网巨头打造的公有云平台，如阿里云、百度云等；第四类为传统的 IT 航母，比如 IBM、SAP、亚马逊等。

(2) 私有云

私有云是为一个客户单独使用而构建的，拥有者之外的用户无法使用。

私有云可由公司自己的 IT 机构，也可由云提供商进行构建。前者一般是在企业或者其他组织在自有的数据中心构建，通常部署在企业数据中心的防火墙内（也可部署在一个安全的主机托管场所），该公司拥有基础设施，并可以控制在此基础设施上部署应用程序的方式；后者采用“托管式专用”模式，由云服务商提供匹配用户需求的云，然后整体租赁给该用户。IBM、惠普、Oracle 等“IT 航母”出于对自有产品的市场推广和深入衍生，一般是私有云的积极推动者。

(3) 混合云

混合云是公有云和私有云两种服务方式的结合（混合），往往是针对一个具体的企业或组织而言的。

企业或组织选择混合云的原因有很多，比如：由于安全和控制原因，并非所有的企业（包括几乎所有的政府机关，如公、检、法、财、税等）信息都能放置在公有云上，所以大部分已经应用云计算的企业都会使用混合云模式。另外，还有很多企业选择混合云的原因是私有云搭建好后，由于业务发展等原因，资源需求量超过了资源池，所以需要通过申请使用公有云作为私有云的补充。比如对一些零售商来说，他们的操作需求会随着假日的到来而剧增，或者是有些业务会有季节性的上扬，此时，先前搭建的私有云无法应付突然爆发的资源需求量。另外一个选择混合云的原因是混合云为企业或组织的弹性需求提供了一个很好的基础，比如灾难恢复。这意味着私有云把公有云作为灾难转移的平台，并在需要的时候使用它，这是一个极具成本效应的理念。

其实还有其他的一些分类标准，比如按云计算面向的行业进行分类，就分为政府云、教育云、金融云、医疗云等，此处不再赘述。

4.4.4 应用提示

事实上,在大数据概念被认可并炙手可热之前,云计算就由工业界发起,然后成为迅速火热起来的一个技术名词或者应用模式,并很快得到包括学术界在内的各界认同。从整体上看,大数据与云计算是相辅相成的,大数据着眼于“数据”,聚焦于具体的业务,关注“数据→价值”的过程,看中的是信息积淀。云计算着眼于“计算”,聚焦于IT解决方案,关注IT基础架构,看中的是计算能力(包括数据处理能力及系统部署能力)。没有大数据的信息积淀,云计算的计算能力再强大也难以找到用武之地,没有云计算的处理能力,大数据的信息积淀再丰富,也难以甚至无法落地。另一方面,云计算涉及的关键技术,如海量数据存储、海量数据管理、分布式计算等也都是大数据的基础支撑技术。

互联网、云计算以及大数据成了三个密不可分的词汇。一般而言,一家互联网公司一定同时是数据公司,不能将数据转换为价值的互联网公司一定不是一个好公司(除非短期内有其他的战略规划)。更进一步,很多公司都可以实现“数据→信息→知识→价值”这样的流程,但是如果不能用最低成本实现上述流程,企业同样难以获益(甚至无法存活),而云计算就是一种低成本实现上述流程的解决方案(之一)。

云计算至少在以下几个环节助力大数据(的落地):

(1) 大数据存储方面

对于任何一个大数据项目的开展,数据规模“大”的特点都会使得存储压力是一个不可小觑的瓶颈,而云计算的存储虚拟化特点几乎保证了存储的无限扩容。同时,通过云计算为多种形态的数据设定统一的接口,这在用户看来,系统所面对的就是形态与分布均统一的数据,即用云计算来屏蔽数据形态与分布的多模性,从而使系统不管是在实现上还是在用户使用时都得到简化。

(2) 大数据计算方面

大数据的核心是获得数据并从数据中提取信息并挖掘知识,任何一个环节都是在计算量特别巨大的同时,算法复杂度(时间复杂度和空间复杂度)也往往极大,如何保障海量计算的吞吐率是大数据项目开展的一个关键。云计算架构的软件系统不仅能提供大规模的数据存储、维护技术,又因其整合了并行计算、分布式计算以及网格计算的特点,拥有强大的分布式以及并行计算的能力,故而能飞速提高算法的运行速度。另一方面,云计算的安全性由于中央集权的数据管理而得到有效提高,“云”使用了数据多副本容错、计算节点同构可互换等措施来保障服务的高可靠性,因此,使用云计算比使用本地计算(机)更可靠。同时,云计算的计算能力虚拟化的特点也使得用户可以透明访问并进行按需扩展。

(3) 大数据部署方面

随着Internet平台及Web开发应用模式的快速发展,基于Web技术开发的分布式应用系统越来越明显地具有软件平台网络化、协作模式联盟化、组织形式虚拟化、业务流程服务化、

资源共享可控化、执行实体协同化等一系列技术特征。而在应用层面上,随着服务计算、云计算等应用模式的普及和推广,基于 Web 技术开发的分布式应用系统的规模也越来越大,应用复杂度越来越高。虚拟技术和弹性可伸缩应用模式的结合,使得云计算能够更好地支持多用户并发访问,并在动态资源分发与监控方面具有更为有效的控制手段,可以高效地支持大规模数据资源的虚拟存储与应用集成。目前,在个性化需求较为突出的应用领域中,为不同用户提供支持个性化服务定制的云计算模式,已然成为一个趋势。

4.5 本章小结

大数据是一个现象,也是人类在不断追求文明的过程中一定会经历的必然阶段,只是在当今信息技术快速发展及人们自身需求不断膨胀的背景下,这个现象突然爆发而已。

每天、每个人都在生产各类数据,比如打电话、发短信、发微博、上微信等,人们在享受社交服务的同时生产了大量的数据,有文字、图片、视频等;还有很多的智能终端,每天在不间断地采集和产生各类数据,比如智能手环采集人的体征数据、气象仪监测大气数据、电子警察探头收集车辆出行数据等;再有各个商家每天产生的大量物流数据、营销数据、客户数据……

人类有别于其他物种的一个特点是从生产实践中发掘知识,并将这种知识以有别于基因遗传的特征传承下来,这也促进了人类自身的不断进步。而上述的各类数据以及根本无法罗列的数据都是人类在各类生产实践中产生的。因而,如何从这些数据中发掘出知识和洞见,是人类必然的一个期望。大数据如此得到关注的一个重要原因就是人们希望能够从这些实践数据里找出知识和洞见,并把它像历史长河中的其他知识一样传递和传承下去。当然,对于有利益诉求的商家而言,更直接的想法是如何从这些数据中发现价值,或许是后者的原因更加触动了人们对大数据的普遍关注、热议和追捧。

从数据到价值的过程一定是“数据→信息→知识→价值”,本章从技术流的角度,从数据的采集、存取、计算架构、数据分析以及服务平台等方面详细阐述了大数据的分析流程与关键技术。应该说,大数据不是技术,而是为了实现大数据关于“数据→价值”的目标,需要一系列的支撑技术。或者,更合理的说法是为了达到这样的目标,需要适合的技术、适合的应用模式、适合的商业模式等共同协作才能最终实现。

晋朝的袁宏在《三国名臣序赞》中提及“形器不存,方寸海纳……”。唐朝的李周翰注曰:“方寸之心,如海之纳百川也,言其包含广也……”

很多人说大数据是数据的海洋,因为不同来源、不同品性的数据如此汇聚在一起恰如不同物质含量的小溪、百川东归大海一样。但同时也意味着,在大数据时代引发的这个“蓝海”中,需要有更多的理论、技术、策略、模式一起投入,才能真正让这个“蓝海”成为真正的价值源泉……

本章参考文献

- [1] Ackroyd S. Data Collection in Context [M]. United Kingdom: Longman Group, 1992.
- [2] Adamic L A, Huberman B A. Power-law Distribution of the World Wide Web [J]. Science, 2000, 287 (5461): 2115-2115.
- [3] Borthakur D, Gray J, Sarma J S, et al. Apache Hadoop Goes Realtime at Facebook [C]. Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, 2011: 1071-1080.
- [4] Dasseni E, Verykios V S, Elmagarmid A K, et al. Hiding Association Rules by Using Confidence and Support [C]. Information Hiding, 2001: 369-383.
- [5] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.
- [6] Gilbert S, Lynch N A. Perspectives on the CAP Theorem [C]. Institute of Electrical and Electronics Engineers, 2012.
- [7] Han J, Haihong E, Le G, et al. Survey on NoSQL Database [C]. Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on. IEEE, 2011: 363-366.
- [8] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques: Concepts and Techniques [M]. Holland: Elsevier, 2011.
- [9] Hoffman S. Apache Flume: Distributed Log Collection for Hadoop [M]. Birmingham: Packt Publishing Ltd, 2015.
- [10] Kimball R, Caserta J. The Data Warehouse ETL Toolkit [M]. Manhattan: John Wiley & Sons, 2004.
- [11] Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection [C]. IJCAI. 1995, 14(2): 1137-1145.
- [12] Peleg D. Distributed Computing [J]. SIAM Monographs on Discrete Mathematics and Applications, 2000.
- [13] Rabkin A, Katz R H. Chukwa: A System for Reliable Large-Scale Log Collection [C]. LISA, 2010, 10: 1-15.
- [14] Rajagopalan R, Varshney P K. Data-aggregation Techniques in Sensor Networks: A Survey [J]. Communications Surveys & Tutorials IEEE, 2006, 8(4): 48-63.
- [15] Shane Cook. CUDA 并行程序设计——GPU 编程指南 [M]. 北京: 机械工业出版社, 2014.
- [16] Surhone L M, Timpelton M T, Marseken S F. Scribe(log server) [M]. Saarbrücken: Betascript Publishing, 2010.
- [17] White T. Hadoop: The Definitive Guide [M]. Sebastopol: O' Reilly Media, Inc., 2012.
- [18] 艾廷华, 成建国. 对空间数据多尺度表达有关问题的思考 [J]. 武汉大学学报(信息科学版), 2005, 30(5): 377-382.
- [19] 关大伟. 数据挖掘中的数据预处理 [D]. 吉林大学, 2006.
- [20] 章沛轩. 一本书读懂大数据 [M]. 北京: 中国商业出版社, 2015.

数据采集与整合

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的刘有力、刘洋、肖雨奇、韩军华、韩建军、汤兆亮、王晓彤等几位同学的协助，在此表示深深的谢意。

5.1 引言

Google 利用其自有平台数据预测流感的爆发趋势及分布。

阿里系利用其自有平台上的交易数据可以预测金融危机的到来。

微博平台利用其自身的数据进行精准广告营销和自动推荐。

银行因为拥有普通民众及企业的交割数据，因此可以对个人和企业进行征信。

工商、税务甚至水、电、煤气公司拥有的数据同样可以作为征信依据……

不同的平台（利益主体）采集和整合的数据除了可服务其自身以外，还可以应用于更高的场景，从而创造出更高的价值（详见 11.2 节）。大数据的核心是从数据中发现知识和洞见，因此，“数据交叉复用”是大数据时代的特征，其中最根本的一点是：“数据为王”。

数据代表着对事物的描述，对数据的记录能力是原始社会与先进社会的一个重要分界标志。公元 3000 年前，在古代埃及、古代中国和印度河流域都发明了记录数据的方法，于是产生了最早期的数据存储方式。人类将在认识和改造自然的过程中发现的自然法则以当时社会特有的数据形式进行记录和存储，这使得人类文明得以延续和发扬。数据存储是数据得以集聚和沉淀的关键，对人类文明的发展具有重要的促进作用。

从世界范围来看，文字作为信息符号诞生后，石头、兽骨等便成为最初的记录载体。商朝后期（公元前 14 世纪—公元前 11 世纪），商朝王室将占卜吉凶的事迹契刻在龟甲或兽骨上，这样便形成了“甲骨文”。随着社会的进步，造纸术的发明和逐渐完善，有力地促进了我国科学文化的传播和发展。古埃及人最早使用的象形文字，经过长期的发展和演变，后来形

成了由字母、音符和词组组成的复合象形文字体系。同时,古埃及人将一种类似芦苇的植物切成长度合适的小段,剖开压平并风干后制成纸草,他们在纸草上写字以记录信息。公元3世纪,古印度的一位科学家巴格达发明了阿拉伯数字,后由阿拉伯人传向欧洲,之后再经欧洲人将其现代化。阿拉伯数字的发明和广泛传播开启了算术的腾飞。

随着社会的不断发展和信息化时代的到来,人类产生数据的形式不仅发生了颠覆式的转变,而且变得越来越多样化,产生的数据量也呈现爆炸式增长。IDC在研究报告《Data Universe Study》中预测全世界的数据量将从2009年的0.8ZB增长到2020年的35ZB,年均增长率大于40%。中国互联网络信息中心(CNNIC)发布的《2013—2014年中国移动互联网调查报告》显示,截至2014年6月,我国手机网民规模为5.27亿,在整体网民中占比达83.4%。移动互联网的迅速发展势必将进一步刺激数据信息的增长和促进大数据时代的发展。

2015年11月11日,阿里巴巴集团网上交易额突破912亿人民币,创造了“24小时内在线零售额最高的单一公司”的吉尼斯世界纪录。如此巨大的交易额背后产生了大量的交易记录 and 商品浏览记录。2014年Facebook每月活跃用户总数已超过22亿,平台上每天产生120亿条消息数据,每天执行的搜索命令达到10亿次,超过150万家企业在Facebook上发布广告信息。2014年Twitter注册用户总量超过10亿,每月活跃用户量为2.55亿,每日发布的微博信息达到3.4亿条。2015年10月,在腾讯全球合作伙伴大会上,微信开发团队公布了相关数据:9月微信用户日登录数量突破5.7亿,日均通话时间2.8亿分钟。2015年8月19日,新浪微博公布的财务报告显示:微博2013年Q4月活跃用户数是1.29亿、日活跃用户数为6140万;2014年Q4月活跃量为1.757亿、日活跃量为8060万;截至2015年Q2月活跃量为2.12亿、日活跃量为9300万;用户数量一直稳步增长……

数据被认为是信息时代的基础生活资料与市场要素,重要程度不亚于物质资产和人力资本。大数据时代,数据逐渐变现为独特的“流通货币”,对数据的掌握与控制将成为业界的新财富。随着企业信息化程度的不断提高,企业内部时刻都在产生着大量的内部数据,如交易记录、系统日志、用户浏览记录等。在大数据时代下,企业数据资源包罗万象,一方面是在与外围客户、合作伙伴通过文本信息、社交网络、移动应用等形式进行互动时产生的大量数据;另一方面是企业内部生产研发、综合办公、视频监控等日常经营管理活动产生的大量信息。互联网的高速发展及移动互联网的迅速普及促进了网络数据的形成。如今,网络数据正在广泛地产生,如电子商务网站中顾客的商品浏览记录、购物清单信息、支付信息,搜索引擎中用户搜索的词条目,新闻网站上发布的新闻动态,社交网站上用户发布的博文信息、用户间的互动信息等。互联网成为大数据使用最广泛、认可度最高的数据源。

大数据的产生是计算机和网络通信技术发展的必然结果,特别是在互联网、云计算、移动互联网、社交网络及物联网等新一代信息技术的发展促进了数据量爆炸式增长的当下。大数据不仅数据量巨大,数据格式也是多种多样,比如存储在关系数据库中的结构化数据、办公文档、XML、系统日志、HTML等半结构化数据,以及图像、文本、语音、视频、电子邮件



等非结构化数据。利用大数据技术分析大数据中隐藏的巨大价值的前提是获取和收集数据本身,因而数据量的大小、数据涉及业务领域的深度,以及数据的质量将对大数据的分析结果产生直接的影响。流行于经济学领域的“二八规则”在大数据应用场景也具有一定的普适性。据统计,大数据价值挖掘过程中有70%~80%的时间花在收集和准备数据阶段,而仅有20%~30%的时间花在数据分析上。如何从深度和广度两个层面进行大数据的采集和整理是本章将要讲述的重点。

本章将尝试梳理大数据项目建设过程中潜在的数据源,以及各类数据源数据的获取方法、策略及技术选型思路,本章后面的结构安排如下:5.2节对大数据来源进行介绍,明确展开大数据采集和整合的潜在数据源,包括内部数据、外部数据和网络数据,并从管理的角度介绍数据质量管控的相关内容;5.3节介绍如何对内部数据源进行数据整合,针对内部数据结构各异的特点讲述ETL数据整合技术和几种常见的开源ETL工具;5.4节介绍如何进行网络数据的抓取,以及针对网络数据分布广、非结构化等特点,介绍网络爬虫技术在网络数据抓取过程中将面临的相关问题;5.5节对本章进行小结。

5.2 大数据的数据源

5.2.1 数据分布

“数据为王”是大数据时代的典型特征,大数据概念的风靡倒逼人们主动思考曾经仅作为成本的数据是否有可发掘的价值,或许正是因为这样的主动发掘,才使得数据的可扩展价值得到了不断的发掘和整合。而如何发掘数据的可扩展价值本身就是一个充满机遇和挑战的命题,需要智慧的人们从不同的角度以不同的思维方式思考,并利用不同的理论和技术手段共同发力、共同拥抱大数据引发的挑战和机遇。当然,这个过程离不开(不同角色的)人们主动、主观的想象力。

作为一个大数据分析师,经常被问及的话题有:

1) 我有这些数据,你能帮我做些什么?

2) 我想做这些事,你需要哪些数据?

3) 有了这些数据,还能做些什么?

……

第一个问题往往是有一定数据基础的甲方(拟开展大数据项目建设的利益主体)在项目伊始,与大数据分析师进行项目研判时经常问及的话题。这个问题也显现了大数据项目开发和传统软件开发的差异,传统软件开发的需求往往是甲方驱动的,而大数据项目的需求是数据驱动的,详细分析请参见10.2.2节。

第二个问题往往是有一定IT建设基础的甲方出于目标产品定位而进行可行性研究时在数据层的慎思。甲方具有这样的思考方式,显然是受到了数据驱动的思维方式的影响。



第三个问题往往是已经以数据驱动的方式开展了一段时间的项目建设后，甲方在数据的原始价值得到实现以后，尝试发掘更多的潜在价值时的思考。这种问题往往是在前段数据驱动的项目建设已经获得了一定的收益（从而让甲方有了更多的价值倾向），或者项目建设前段时间已经为数据采集投入了相当多的资金（甲方从建设角度需要尽快收回成本）的前提下提出的，这就意味着，发掘数据的可扩展价值往往尤为重要，这是保证甲方对项目开展抱有持续信心的重要基础。

类似的问题事实上还有很多，而且彼此都是耦合的，归结为在“数据→价值”的价值实现过程中，具有绝对主观性的价值目标在不断迭代（更新、更改、膨胀等）的过程中，在数据层应该有的响应及在数据需求得到满足后，在价值目标方面能够进行的继续迭代。本章将重点关注：围绕具体的价值目标，需要什么样的数据，这些数据存放在何处？而对于如何实现“数据→价值”的具体细节将在后文第10章中详细介绍。

对于一个计划进行大数据项目建设的利益主体而言，可以考虑的数据源如表5-1所示。

表 5-1 数据来源分布

数据来源	数据类别	描述
本单位自营	自营系统（平台）	本单位自营，理论上数据可以最大限度共享
	历史遗留数据	纸质文档或存放在历史数据库中
	其他利益主体运营平台	与自营系统（平台）类似，仅归属不一样
外单位他营	物联网数据（源）	一般有具体利益主体运营，与自营系统（平台）类似，仅归属不一样
	政府数据	政府出于社会监管需求而营建的各个运营平台，一般是归属于各个政府部门，并且是存放于各个子系统服务器中，或者散布于互联网中
	互联网/移动互联网数据	主要以网页的形式存放于互联网中（实际也是存放于不同利益主体的服务器上）

需要说明的是：

1) 开展一个大数据项目，其数据源一定来源于本单位自营（数据主权归属于本单位）和外单位他营（数据主权不归本单位所有）两个方面。

2) 互联网数据是一种比较特殊的外单位他营数据，在某种意义上而言，这些数据也是存放在其他利益主体的服务器上的，不过基于互联网的共享精神，所有人都可以通过网络访问（浏览网页）的形式获得相关数据。另一方面，互联网数据几乎沉淀了所有人、事、物的大多信息，因此，从互联网上采集和整合相关数据几乎是所有大数据项目建设中的必然途径。

3) 政府数据是政府出于社会监管等原因而架设的各个应用系统收集、整理和使用的各种数据，这些数据往往具有较高的真实性、权威性和实时性，因此，政府数据的采集与整合也几乎是所有大数据项目建设的重要数据渠道。

4) 物联网数据应该是大数据的重要来源，不过，从数据源的角度而言，物联网数据往往存放于各个利益主体的服务器上，不能像互联网数据那样允许其他人自由地访问，因此针对

物联网数据的采集,往往需要与当事人的利益主体进行商务洽谈和合作,用类似于其他利益主体运营平台的数据采集方式进行整合。

从技术流的角度看,从互联网上采集数据利用的是爬虫技术。而从其他数据源,包括本单位自营系统、外单位利益主体营运的系统、政府数据等,本质上都是直接在数据库层面或软件应用层面进行的数据交换,两者的技术思路截然不同。因此,可以从技术流上将数据的来源划分为内部数据(散布于各个利益主体,包括政府各级部门及企事业单位的服务器中,是在数据库层面或软件系统层面进行的数据导入导出)和互联网数据(散布于互联网中,也称为网络大数据,通过网络爬虫自动从 URL 中获得数据)。

5.2.2 内部数据

本节所提及的“内部数据”是数据采集与整合技术流上的一个分类,专门指那些不同的利益主体(包括政府各个部门、企事业单位等)出于自身职能定位和获益诉求而建设的 IT 系统在完成本部门既定角色目标任务的过程中,有意或无意地存储下的有关物理世界实体对象的各类数据。具体而言,这些数据的分布情况如下(包括但不限于):

1. 政府数据

政府出于社会管理的目的而下设的各种部门,比如公检法、财政部、发改委、工商、税务、海关、人社、医疗等,所有这些组织机构出于有效完成部门职能的目的,会构建很多业务系统,这些业务系统产生的数据主要以特定的结构存储在相应的数据中心。这些数据内部蕴含着巨大的价值,能够为政府宏观政策的制定、国家安全防护、社会有效管理等提供有力的数据支撑。这些政府数据往往具有可信度高、完整性好、实时性强、实体对象描述指向性明确且具体等特点,因此,在进行大数据项目的建设过程中,通过某种渠道采集相关政府部门的数据,已经成为一个必然的趋势。不过政府数据的采集也存在很多的挑战和困难,比如(包括但不限于):

1) 出于数据安全及涉密的考虑,以及制度的规定,政府数据往往具有很强的封闭性(开放性弱),这使得政府数据的获取成本往往极高(包括商务成本、技术成本和制度规避的成本)。

2) 根据不同的职能定位,不同政府部门运营和管理的数据往往仅与该部门独立职能相关,因此,每一个部门的政府数据都缺乏(一定意义上的)全局性,这就意味着采集更为全面的政府数据代价极大。

3) 各级政府部门的信息基础设施建设不均衡,这使得相同类型的数据在不同级别的政府部门的服务器上表现形式会不完全一样,这也给数据的采集与整合带来极大的困难。

2015年9月5日,国务院发布了《国务院关于印发促进大数据发展行动纲要的通知》,为政府各大部门的数据开发及共享设定了时间表,这对于大数据项目的建设而言,是极大的福音。

2. 各利益主体自营数据

各利益主体（事实上包括政府各级部门、企事业单位等）出于不同的获益需求，会构建不同目标应用的 IT 系统，比如 ERP、在线办公、在线交易等，这些系统在有效完成各个单位的主营业务的同时，汇聚了大量相关数据，这些数据以本单位私有财产的形式存放在各自的服务器中。显然，这些数据在辅助实现各个业务系统的价值目标方面具有重要的意义。同时，这些数据也为各个利益主体的商务智能的实现提供了重要的数据基础保障。

应当注意到，随着大数据时代的来临，各个利益主体在数据的使用和营运方面也出现了很大的变化。其中一个重要的变化在于：互联网的不断发展和对各个领域的渗透，使得各个利益主体开始有意识地将互联网作为一个工具、渠道或平台，将自有 IT 系统从不同的层次和角度，通过改良，然后嫁接到互联网之上，从而实现更好的产品设计、制造和营销等。所有这些动作都在逐步淡化各个单位自有内部数据和互联网数据的界限。也就是说：单位内部的信息化应用环境在不断发生变化，传统意义上的互联网数据正从外部数据逐步被纳入本单位的内部数据管理。

如何有效利用这些已经泛化的“内部数据”并实现精细化管理，已然成为任何一个利益主体的共同需求。特别是在利益主体营建大数据项目时，如何对这些数据进行有效的集成和汇聚，是大数据项目建设的一个重要基础，不过各个利益主体的数据采集也存在着很多的挑战和困难，比如（包括但不限于）：

1) 不同的利益主体所拥有的数据在目标应用中的价值度是不一样的，往往各个利益主体的数据仅仅反映了某一个维度的价值趋势，而如何选择更多的彼此互补的数据源（合作单位）本身就是一个难题，这涉及不同利益主体的数据评估问题，是一个技术问题，同时还受大数据项目建设的物理条件的约束。

2) 在采集和整合不同利益主体的数据，从而为目标应用提供数据支撑时，一个非技术因素的商务难题在于潜在合作单位是否愿意将数据共享。涉及的商务问题包括：对方是否有合作的意愿，以及在有意愿的前提下如何进行有效的合作。

3) 不同利益主体的信息基础设施建设不均衡，这使得相同类型的数据在不同利益主体的服务器上的表现形式不完全一样，这给数据的采集与整合带来了极大的困难。

3. 物联网数据

物联网快速发展的同时也制造了海量数据，如何妥善处理及合理利用这些海量数据是物联网下一步发展的关键。当然，物联网所产生的数据本身的复杂性势必加大了物联网领域大数据落地的难度，此处不再赘述。

本节关注的是，在大数据项目建设中，将物联网数据的收集纳入数据采集的考虑范畴也是一种必然的趋势。根据物联网终端或相应 APP 建设单位的不同，物联网数据或者以企业自营数据库的形式存放于企业内部的数据库中，或者以开放共享的精神存放于互联网中。因为前者的数据采集类似于上文提到的各利益主体自营数据的采集，因此此处不再赘述；对于存

放在互联网中的数据,其具体的数据采集方式可参见后文,此处也不再赘述。

5.2.3 互联网数据

本节所提及的“互联网数据”是数据采集与整合技术流上的一个分类,专门指那些通过不同的互联网应用产品而沉淀在互联网中的各类数据。事实上,这些数据也都存放在不同利益主体的服务器中,不过由于互联网的开放、共享精神,普通人都可以通过浏览网页(或者通过APP)的形式访问这些数据。具体而言,这些数据的分布情况如下(包括但不限于):

1) 门户网站出于其媒体属性所发布的新闻、评论、报道等,比如新浪财经、搜狐新闻等,这些数据往往具有较强的实时性和专业性。

2) 政府部门出于信息公开的目的在互联网上公开的数据,比如法院公告、工商缺陷产品召回信息、政府招标信息等,这些数据往往具有很高的权威性和可信性。

3) 社交网站出于其媒体属性和社会属性允许普通用户发表自媒体信息,在提供用户社交服务的同时,将用户的言论、生活轨迹等记录下来,这些数据往往具有一定的实时性和针对性。

4) 电商网站出于营销目的允许用户自由采购产品并查询、发布产品评论及销售量信息,这些数据往往具有一定的真实性和实时性。

5) 论坛(含地方论坛)往往是网民发表意见舆情的开放渠道和平台,用户在发表个人意见的同时,自己的价值倾向、事件评估等信息也被网站记录了下来,这些数据往往具有一定的实时性和针对性。

还有大量的其他类型的互联网数据,此处不再一一罗列。互联网数据中沉淀着大量能反映用户偏好倾向、事件趋势等的相关信息。更重要的是,互联网数据均是以共享和开放的精神存放于互联网中的,这就意味着进行互联网数据采集的成本往往较低,因此对相关目标应用而言,进行相关互联网数据的采集和整合几乎成为大数据项目建设的必然选择。不过进行互联网数据的采集也存在很大的困难和挑战,具体有(包括但不限于):

1) 各个网站的IT建设水平不一样,以及出于不同的用户体验,各个网站的模板结构往往也不一样,这就意味着,通过统一的方法从互联网中采集数据几乎是不可能的。针对不同的网站开发不同的采集方法,其困难也不容小觑。

2) 从互联网中获取数据都是通过爬虫程序自动进行的(见后文5.4节),不同的网站出于对爬虫程序的监管(其实,每个网站既希望有更多的访问流量,又担心爬虫程序的访问影响网站的服务质量),往往会设置很多障碍(比如验证码),从而增加互联网数据采集的难度。

3) 互联网数据往往是以文本、表格、图片、视频等形式存在,这也给互联网数据的采集带来了挑战和困难。

5.2.4 应用提示

数据采集和整合是大数据项目的基础,以数据的归属单位来看,大数据项目的数据一方

面来源于本单位的自营数据,另一方面来源于外单位的他营数据;从技术手段上看,数据一方面通过数据交换,在数据库层次或软件应用层次实现其他 IT 系统的数据导入导出,另一方面,通过网络爬虫从互联网中采集和整合。在实际应用中,需要注意以下几点:

1) 本单位自营数据往往相对容易采集和整合(技术流上有要求,此处不再赘述),操作上的难题是:一个单位物理上往往会划分为很多部门,各个部门之间存在内部利益冲突或审计制度等问题,进而导致数据采集和整合会因为非技术因素而无法开展。在涉及政府各个部门之间的利益协调上,这种情况尤为突出。

2) 外单位他营数据(包括其他利益主体的业务系统、政府各个部门的业务系统、物联网应用等)的采集往往需要一定的商务支撑,比如通过购买或某种利益的交换。但究竟与哪些外单位进行合作,这需要大数据项目的建设方、承建方、用户和开发人员等多边研判才能决定。

3) 互联网数据是一种特殊的外单位自营数据,从互联网上采集和整合相关数据几乎是所有大数据项目建设中的必然途径,但究竟从哪些网站及 URL 中获取数据,这需要大数据项目的建设方、承建方、用户和开发人员根据目标应用及价值取向等进行多边研判。

4) 在进行内部数据及互联网数据的集成时,应当注意到互联网数据在实时性方面往往具有较为明显的优势,但是数据的真实性及数据的质量往往远劣于其他数据来源。因此,在实际数据集成(应用)过程中,需要根据不同的分析目标运用不同的策略来应对。

5) 数据的质量是大数据项目建设的重要基础。一般而言,数据质量的要素包括数据的准确性、完整性、适用性、实时性、有效性。这意味着,在数据采集伊始,就要建立匹配的数据质量监管制度,并在整个大数据项目的建设过程中,引入数据质量监管技术对数据整个生命周期进行监管。

5.3 内部数据及内部数据采集

5.3.1 目标任务

对一家企业来说,企业数据不仅包括本企业自己运营所产生的数据,还包括该企业与其他企业合作时可以获得的数据:

1) 在企业内部,组织经营、管理和服务等业务流程中产生了大量的数据并被存储于数据中心或数据集中。这些数据虽然都是由同一企业的内部业务所产生,但一般是由不同的系统产生并以不同的数据结构存储在不同的数据库中。如 ERP 系统产生的数据存储在 ERP 数据库中,在线交易平台产生的数据存储在交易数据库中。在企业内部的信息化系统中,也会产生一定的半结构化数据,如交易日志、用户浏览日志,以及各种监控设备所产生的视频、音频等非结构化数据。

2) 另一方面,在企业运营过程中可能会涉及其他合作企业的数据,这些数据可能是以合

作企业通过数据推送的方式提供,也有可能是通过数据接口访问的方式提供,或者由合作企业直接提供数据库访问权限的方式提供。从企业数据的组成可以看出,企业数据的数据来源较多、数据组织形式多样,因此整合企业数据是一项具有挑战性的任务。

面对大数据时代带来的机遇和挑战,内部数据资源整合是现代企业必须具备的能力和强有力的竞争优势,体现在如下几个方面:

1) 构建数据驱动应用,推进扩展价值实现:以自营数据为中心,围绕本单位的既有价值定位,探索数据驱动的创新应用研发,以此拓展数据的可扩展价值。

2) 统一数据规范标准,推动数据共享开放:以自营数据为中心,围绕本单位的既有业务逻辑,探索数据标准及接口规范,从而为数据的开放共享提供支撑。

3) 重视数据安全,完善数据安全保障:以数据安全为前提,与上下游企业,以及安全管理机构、评测机构等第三方机构开展广泛合作,从企业管理制度、流程和技术手段等多方面确保大数据生态圈的数据安全。

4) 推进数据融合管理,增加数据语义厚度:推进结构化和非结构化数据的融合式发展,将超文本、超媒体数据模型与面向对象数据模型进行融合,构建适合结构化和非结构化数据统一组织和管理的数据库模型,为数据的有效利用提供支撑。

由于不同用户和企业内部不同部门提供的内部数据可能来自不同的途径,其数据内容、数据格式和数据质量千差万别,有时甚至会遇到数据格式不能转换或转换数据格式后丢失信息等棘手问题,严重阻碍了各部门和各应用系统中数据的流动与共享。因此,能否对数据进行有效的整合将成为是否能够对内部数据进行有效利用的关键,ETL是整合数据的一个重要的技术手段。

5.3.2 关键技术

ETL即数据抽取(Extract)、转换(Transform)、加载(Load)的过程。ETL是将企业内部的各种形式和来源的数据经过抽取、清洗转换之后进行格式化的过程。ETL的目的是整合企业中分散、零乱、标准不统一的数据,以便于后续的分析、处理和运用。

一个简单的ETL体系结构如图5-1所示。

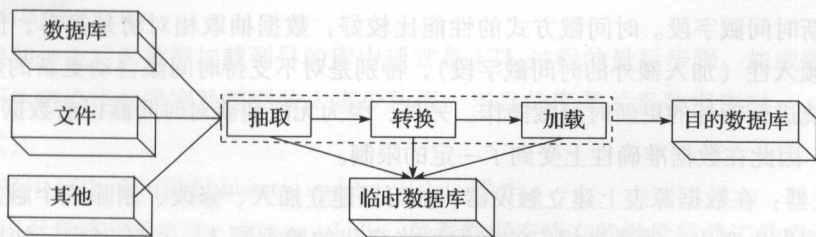


图 5-1 ETL 系统结构图

ETL过程中的主要环节包括数据抽取、数据转换和加工、数据加载。为了实现这些功能,ETL相关工具一般会进行一些功能上的扩充,如 workflow、调度引擎、规则引擎、脚本支持、

统计信息等。

1. 数据抽取

数据抽取是从数据源中抽取数据的过程。由于在大多数场景下，数据会存放在数据库里，数据抽取也就变成了从数据库中抽取数据的过程。从数据库中抽取数据一般分为两种方式：全量抽取和增量抽取。

(1) 全量抽取

全量抽取就是对整个数据库的所有数据进行抽取，它将数据源库中的所有数据原封不动地从数据库中抽取出来，然后转换成自己的 ETL 工具可以识别的格式。由于全量抽取是对整个数据库的所有数据进行抽取，不需要进行其他复杂处理，因此抽取过程比较直观、简单。但是在实际运用中，很少会用到全量抽取，主要是因为数据是实时增加的，全量抽取在每次抽取的时候将会重复抽取上次已经抽取过的历史数据，这样在产生大量冗余数据的同时也降低了抽取的效率。于是增量抽取策略被广泛关注，并得到广泛应用。

(2) 增量抽取

增量抽取只抽取自上次抽取以来数据库中要抽取的表中新增或修改的数据。如何捕获变化的数据是增量抽取的关键。优秀的捕获方法应该能够做到以较高的准确率获得数据库中发生变化的数据，同时还不能对业务系统造成太大的压力而影响现有的业务。在增量数据抽取的过程中，常用的捕获变化数据的方法有：日志比对、时间戳、触发器、全表比对等。

1) 日志比对：通过分析数据库自身的日志来判断发生变化的数据。以常用的 Oracle 数据库为例，Oracle 数据库具有改变数据捕获 (Changed Data Capture, CDC) 的特性。CDC 能够帮助用户识别从上次抽取之后发生变化的数据。利用 CDC 在对源表进行 insert、update 或 delete 等操作的同时就可以提取数据，并且将变化的数据保存在数据库的变化表中。这样就可以捕获发生变化的数据，然后利用数据库视图以一种可控的方式提供给目标系统。

2) 时间戳：通过增加一个时间戳字段，在更新修改表数据的同时修改时间戳字段的值。当进行数据抽取时，通过比较系统时间与时间戳字段的值来决定抽取哪些数据。对于支持时间戳自动更新的数据库来说，在数据库表其他字段的数据发生改变时，系统将自动更新时间戳字段的值。有的数据库不支持时间戳的自动更新，这就要求业务系统在更新业务数据的同时，手工更新时间戳字段。时间戳方式的性能比较好，数据抽取相对清楚简单，但对业务系统有很大的侵入性 (加入额外的时间戳字段)，特别是对不支持时间戳自动更新的数据库，还要求业务系统进行额外的更新时间戳操作。另外，因为无法捕获对时间戳以前数据的 delete 和 update 操作，因此在数据准确性上受到了一定的限制。

3) 触发器：在数据源表上建立触发器，如可以建立插入、修改、删除三个触发器，每当源表中的数据发生变化，就通过相应的触发器将变化的数据写入一个临时表，抽取线程从临时表中抽取数据，临时表中抽取过的数据被标记或删除。触发器方式的优点是数据抽取的性能较高，缺点是要求业务表建立触发器，对业务系统有一定的影响。

4) 全表比对：典型的全表比对的方式是采用 MD5 校验码。ETL 工具事先为要抽取的表建

立一个结构类似的 MD5 临时表, 该临时表记录源表的主键及根据所有字段的数据计算出来的 MD5 校验码。每次进行数据抽取时, 对源表和 MD5 临时表进行 MD5 校验码的比对, 从而决定源表中的数据是新增、修改还是删除, 同时更新 MD5 校验码。MD5 方式的优点是对源系统的侵入性较小 (仅需要建立一个 MD5 临时表), 其缺点也是显而易见的。与触发器和时间戳方式中的主动通知不同, MD5 方式是被动地进行全表数据的比对, 性能较差。当表中没有主键或唯一列且含有重复记录时, MD5 方式的准确性就较差。

ETL 处理的数据源除了关系数据库外, 也可以是文件, 如 TXT 文件、EXCEL 文件、XML 文件等。对文件数据的抽取一般是进行全量抽取, 每次抽取前可保存文件的时间戳或计算文件的 MD5 校验码, 下次抽取时进行比对, 如果相同则可忽略本次抽取。

2. 数据转换和加工

从数据源中抽取的数据不一定完全满足目的数据库的要求, 如数据格式的不一致、数据输入错误、数据不完整等, 因此有必要对抽取出来的数据进行数据转换和加工。数据的转换和加工可以在 ETL 引擎中进行, 也可以在数据抽取的过程中利用关系数据库的特性同时进行。

(1) ETL 引擎中的数据转换和加工

在 ETL 引擎中一般以组件化的方式实现数据转换。常用的数据转换组件有字段映射、数据过滤、数据清洗、数据替换、数据计算、数据验证、数据加解密、数据合并、数据拆分等。这些组件如同一条流水线上的一道道工序, 它们是可插拔的, 且可以任意组装, 各组件之间通过数据总线共享数据。有些 ETL 工具还提供了脚本支持, 使得用户可以以一种编程的方式定制数据的转换和加工行为。

(2) 在数据库中进行数据加工

关系数据库本身已经提供了强大的 SQL 指令和函数来支持数据的加工, 如在 SQL 查询语句中添加 where 条件进行过滤、查询中重命名字段名与目的表进行映射、substr 函数、case 条件判断等。相比在 ETL 引擎中进行数据转换和加工, 直接在 SQL 语句中进行转换和加工更加简单清晰, 性能更高。但是有些应用过程比较特殊, 使得 SQL 语句无法处理, 这时可以交由 ETL 引擎进行处理。

3. 数据加载

将转换和加工后的数据加载到目的库中通常是 ETL 过程的最后步骤。加载数据的最佳方法取决于所执行操作的类型及需要装入多少数据。当目的库是关系数据库时, 一般来说有两种加载方式:

- 1) 直接用 SQL 语句进行 insert、update、delete 操作。

- 2) 采用批量装载的方法, 如 bcp、bulk、关系数据库特有的批量装载工具或 API。

大多数情况下会使用第一种方法, 因为它们进行了日志记录并且是可恢复的, 而批量装载操作易于使用, 并且在装入大量数据时效率较高。具体使用哪种数据装载方法取决于业务系统的需要。

5.3.3 ETL 工具

在大数据时代,数据的异构性使直接利用数据展开分析变得很难,因此进行异构数据整合是数据分析的前提,其结果的好坏将影响后续工作的准确性。ETL 过程是进行异构大数据整合的必备过程,许多公司展开 ETL 工具的开发,目前市面上的 ETL 工具众多: Informatica PowerCenter(Informatica 公司开发)、DataStage(Ascential 公司开发,2005 年被 IBM 收购)、Kettle(业界最有名的开源 ETL 工具)、ETL Automation(NCR Teradata 公司开发)、OWB(Oracle Warehouse Builder)、ODI(Oracle Data Integrator)、Data Integrator(Business Objects 公司开发)、DecisionStream(Cognos 公司开发)等。本节将重点介绍几个常用的 ETL 工具。

(1) DataStage

DataStage 是 IBM 公司推出的一套集成工具,专门用于对多种数据源的数据抽取、转换和维护过程进行简化和自动化,并将其输入数据集市或数据仓库等目标数据库。DataStage 能够处理多种数据源的数据,包括主机系统的大型数据库、开放系统上的关系数据库和普通的文件系统等。

关于 DataStage 的详细介绍参见官网 <http://www-03.ibm.com/software/products/en/ibminfo-data>。

(2) Informatica PowerCenter

Informatica PowerCenter 是 Informatica 公司开发的为满足企业级的要求而设计的企业数据集成平台,并提供企业部门的数据和电子商务数据源之间的集成,如 XML、网站日志、关系型数据、主机和遗留系统等数据源。

官网 <https://www.informatica.com/products/data-integration/powercenter.html> 给出了有关 Informatica PowerCenter 的详细介绍,更多详情请参见官网。

(3) Kettle

Kettle 是一款由 Java 编写的开源的 ETL 工具,可以在 Windows、Linux、UNIX 上运行,数据抽取高效稳定。Kettle 工具集包含 4 个产品——Spoon、Pan、Chef、Kitchen,具体如下:

- Spoon 是转换设计工具,使用户通过图像界面来设计 ETL 的转换过程。
- Pan 是转换执行器,在后台批量运行由 Spoon 设计的 ETL 转换过程。
- Chef 是任务设计器,允许用户创建新任务。
- Kitchen 是任务执行器,批量执行由 Chef 设计的任务。

Kettle 中有两种脚本: transformation 和 job,其中 transformation 完成对数据的基础转换,job 则完成整个工作流的控制。

关于 Kettle 的详细介绍详参见其官网 <http://community.pentaho.com/projects/data-integration/>。

表 5-2 给出了上述 3 种主流 ETL 工具的特点与区别,仅供参考。

表 5-2 3 种主流 ETL 工具的特点与区别

比较维度	DataStage	Informatica PowerCenter	Kettle
数据源	目前市场上的大部分主流数据库, 并且具有优秀的文本文件和 XML 文件读取和处理能力	大部分主流数据库, 用于访问和集成几乎任何业务系统、任何格式的数据	大部分主流数据库
免费与否	需购买	需购买	免费开源
运行平台	Windows/UNIX/Linux	Windows/UNIX/Linux	Windows/UNIX/Linux
软件安装和升级	图形安装, 安装步骤较为复杂	完全图形化安装, 无须额外安装平台软件, 且无须修改系统内核参数	绿色安装, 直接使用
处理性能	支持并行处理, 此外 DataStage 企业版可以在多台装有 DataStage Server 的机器上并行执行。并行执行能力使得 DataStage 所能处理数据的速度可以得到趋近于线性的扩展, 可以轻松处理大量数据	可并行运行多个 Session 以提高性能, 可使用分区写目标数据以提高速度, 可建立多个 PowerCenter Server, 并发运行多个 Session 和 workflow。结合 Streaming 和文件交换区的技术, 优化硬盘和内存的资源利用。Session 支持多线程和管道技术 (Pipeline)	使用 JDBC, 性能与 DataStage、Informatica 相比要差很多, 适合于数据量较小的 ETL 加工使用
元数据管理	元数据信息不公开	元数据资料库可基于所有主流系统平台的关系型数据库 (Oracle、DB2、teradata、Informix、SQL Server 等)	无元数据管理
抽取容错性	没有真正的恢复机制	抽取出错可恢复, 可实现断点续传的功能	无恢复功能
操作便捷性	全图化开发, 无编码	全图化开发, 无编码, 操作简便	全图化开发, 无编码, 操作简单
编码支持	几乎支持目前所有的编码格式	支持编码格式十分丰富	支持常见的编码格式
系统安全性	只提供 Developer 和 Operator 两个角色, 系统较安全	多范围的用户角色和操作权限 (只读、操作和设计等), 权限可以分到用户或组	简单的用户管理功能

5.3.4 应用提示

在大数据项目建设过程中, 数据采集是最关键的环节。数据采集的数据源对象包括本单位的自营数据、其他单位的他营数据、互联网数据, 本节提到的内部数据是相对于后文的互联网数据而言的。这里隐含的一个问题是: 除了本单位的自营数据以外, 还应该采集哪些其他单位的他营数据, 这个问题与目标定位有关 (往往需要开发方、用户方等多边研判后加以遴选), 也与商务合作策略及进度有关。在有商务运维的支撑下, 纯粹的数据采集涉及的相关应用提示如下 (包括但不限于):

1) 在系统初始上线前, 将既有数据导入新系统中, 通常是不可或缺的步骤。如果数据是

在既有自营系统（平台）中，则需要在新系统上线伊始，就从既有自营系统（平台）中将数据（批量）导出输入到新系统中；待系统正常有序地运行后，还需要配置一定的策略（比如频度、数据交换时间等）与既有自营系统（平台）进行（增量）数据交换。

2) 在实际操作中，往往还有一类既有数据，它是以历史文档的形式保存在甲方的数据库里，甚至是以纸质（非电子存储）的方式存放在甲方的档案馆里。这类数据往往也需要纳入数据导入的范围。特别是对于非电子存储的数据，意味着首先需要数字化，然后进行数据的导入，这种应用场景在很多传统行业和领域极其常见。显然，这一步骤的工作量和成本都极其巨大，这也是本书在 3.4.2 节中提及的大数据项目建设应该在有充分 IT 建设基础的场景下优先进行的原因。

3) ETL 是在数据库层进行数据交换的一个工具。这意味着，需要对数据源侧的数据字典、数据组织与表结构有清晰的了解，否则极易出现数据获取不完整的情况。在实际进行 ETL 的过程中，新系统的开发者往往缺乏既有系统（数据源侧）的技术支撑，这通常会给新系统研发（在数据交换方面）带来很大的困难。

4) 传统意义上，ETL 的流程是先抽取（从原始数据库中提取出数据）、再转换（转换成目标数据库的格式）、最后加载（将转换好的数据导入目标数据库）。在大数据场景下，出于对数据加载效率的考虑，一般将顺序更改为 ELT，即先抽取、再加载、最后转换，这样做的最大动机在于先将原始数据库的数据最大范围地导入当前数据库中（提升加载效率），至于如何处理和转换则放在之后进行。

5) ETL 只是众多数据交换方式中的一种，非常适合于大批量的数据导入导出的场景（比如系统上线伊始），在系统正常运行的状态下，当然也可以通过 ETL 工具在后台完成数据的交换（按照某种策略），事实上，还有一种 API 接口方式（由原系统提供类似功能，新开发系统按照 API 规约进行数据存取）也可以进行数据的交换，这种情况特别适合于系统在线运行状况下进行实时（新增）数据的导入导出。

6) 在实际应用场景下，API 接口方式非常适合于本单位与外单位进行数据交换的场合，这种数据交换的场合往往是建立在一定的商务模式的基础上，双方达成数据交换的意向，数据源侧出于对数据库访问安全的控制，往往不会倾向于在数据库层进行 ETL，而 API 接口方式则是一种非常好的补充。

7) 成熟的 ETL 产品及开源的 ETL 产品有很多，5.3.3 节仅给出了一些示例，在实际操作中，可以根据项目开发的成本限制及团队的技术研发水平，在更广泛的范围内进行技术选型和产品选型。

8) 另外一点，前文提及的 ETL 或 API 都是在说新系统通过怎样的手段获得原系统的数据，而事实上，任何一个新开发的大数据系统平台，本身都应该兼具的一个隐含功能是：将本系统的数据以服务的形式提供给第三方。这就意味着，在进行新系统研发的过程中，系统设计者应该有意识地设计与实现面向第三方数据访问的 API 接口方式，允许第三方获得当前系统的数据。

5.4 互联网数据及互联网数据采集

5.4.1 目标任务

近年来,包括互联网、物联网、云计算等信息技术在内的IT通信业迅速发展,促使现代信息社会步入了大数据时代。数据的快速增长给许多行业带来了严峻挑战的同时,也为企业的快速发展提供了宝贵的机遇。电子商务、互联网金融和社交网络等新型行业的兴起及飞速的发展,在极大地改变人们的生活、购物及交流方式的同时,也产生了大量的网络数据,如交易数据、博文图片信息、地理位置信息等。2015年全国两会期间,李克强总理在政府工作报告中提出了“互联网+”战略,这在加快互联网与传统行业融合的同时,也进一步丰富了网络大数据的来源。

2015年3月5日十二届全国人大三次会议上,李克强总理在政府工作报告中首次提出“互联网+”行动计划。“互联网+”是对“互联网改造传统产业”模式的进一步深入和发展,它是创新2.0下互联网发展的新形态、新业态,是知识社会创新2.0推动下的互联网形态演进。伴随知识社会的来临,驱动当今社会变革的不仅仅是无所不在的网络,还有无所不在的计算、数据和知识。“互联网+”不仅仅让互联网“移动了”、“泛在了”、应用于某个传统行业了,更加入了无所不在的计算、数据和知识,造就了无所不在的创新,推动了知识社会以用户创新、开放创新、大众创新、协同创新为特点的创新2.0,改变着我们的生产、工作、生活方式,也引领了创新驱动发展的“新常态”。

网络大数据(Network Big Data)通常是指“人、机、物”三元世界在网络空间中彼此之间相互交互与融合所产生的并在互联网上可获得的大数据。网络大数据不仅数据量级大,而且具有一些其他数据源所不具备的特性:

- 1) 多源异构性:网络大数据通常由不同的用户、不同的网站所产生,数据形式也呈现出不同的形式,如语音、视频、图片、文本等。
- 2) 交互性:不同于测量和传感器获取的大规模科学数据(如气象数据、卫星遥感数据),微博、微信、Facebook、Twitter等社交网络的兴起导致大量网络数据具有很强的交互性。
- 3) 时效性:在互联网和移动互联网平台上,每时每刻都有大量的新数据发布,网络大数据内容不断发生变化,使得信息传播具有时序相关性。
- 4) 社会性:网络上用户不仅可以根据需要发布信息,也可以根据自己的喜好回复或转发信息,网络大数据直接反映了社会状态。
- 5) 突发性:有些信息在传播过程中会在短时间内引起大量新的网络数据的产生,并使相关的网络用户形成网络群体,体现出网络大数据及网络群体的突发特性。
- 6) 高噪声:网络大数据来自于众多不同的网络用户,具有很高的噪声和不确定性。

的网页信息。网络爬虫抓取网页的流程如下：

- 1) 指定入口 URL，将其加入种子 URL 队列中。
- 2) 将种子 URL 加入待抓取 URL 队列中。
- 3) 从待抓取 URL 队列依次读取 URL，从互联网中下载 URL 所链接的网页。
- 4) 将网页的 URL 保存到已抓取 URL 队列中，将网页信息保存到下载网页库中。从网页中抽取出需要抓取的新 URL 并加入待抓取 URL 队列中。
- 5) 持续上述步骤 1 至 4 直到待抓取 URL 队列为空。

根据不同的应用，爬虫系统在许多方面都存在着差异，按照网络爬虫的功能可以将其分为批量型爬虫、增量型爬虫和垂直型爬虫 3 类，它们之间的具体区别与联系参见表 5-3。

表 5-3 3 类典型爬虫

爬虫类别	功能描述	适用场合
批量型爬虫	根据用户配置进行网络数据的抓取，此处的用户配置包括： 1) URL 或 URL 列表（往往也称为 URL 池） 2) 爬虫累计工作时间 3) 爬虫累计获取的数据量 4) 其他	1) 互联网数据获取的任何场合，往往用于评估算法是否可行及审计目标 URL 数据是否可用 2) 批量型爬虫事实上是另外两类爬虫的基础
增量型爬虫	根据用户配置持续进行网络数据的抓取，此处的用户配置包括： 1) URL 或 URL 列表（往往也称为 URL 池） 2) 单个 URL 数据抓取频度 3) 数据更新策略 4) 其他	（准）实时获取互联网数据的任何应用场景（通用的商业搜索引擎爬虫基本都属此类）
垂直型爬虫	根据用户配置持续进行指定网络数据的抓取，此处的用户配置包括： 1) URL 或 URL 列表（往往也称为 URL 池） 2) 敏感热词 3) 数据更新策略 4) 其他	（准）实时获取互联网中与指定内容（一般通过配置 URL 池或热词的方式设定）相关的数据（垂直搜索网站或垂直行业网站往往需要此种类型的爬虫）

需要补充说明的是：

1) 在实际操作中，我们往往会设定一个有限数量的 URL 列表（尽管数量或许会很大），然后让爬虫从这些指定的 URL 池中按照某种策略顺序（或并行）地获取数据。

2) 在实际操作中，URL 列表的设置（与维护）是一个经验性很强的工作，需要对目标应用场景有极强的敏锐度，往往由领域用户（或专家）协同研判来进行。

3) 在实际操作中，任何一个 URL 能否被纳入正式的 URL 池是需要经过“假想—验证”的流程来评估的，经过评估认为可行且可信的 URL 才会被纳入 URL 池中。具体过程是：①用户假定某 URL。②利用上述的批量型爬虫从这个 URL 中获取数据。③将从此 URL 抓取的数据交由用户评估，用户从数据可行性（是否对目标有用）、技术可行性（爬虫是否能够完备可信地获取数据）等角度进行评估。④经过评估认为此 URL 数据有用且有效，则将此 URL 正式配

置进入 URL 池。当然这个“假想—验证”的过程也用于作为评估爬虫程序本身是否支持当前 URL 数据的获取的重要依据，也是爬虫程序改进的重要驱动源。

不论是哪一种类型的爬虫，其执行步骤均是：从 URL 池中选择一个具体的 URL，然后利用爬虫，从这个 URL 中获取数据。不过需要注意的是，每一个 URL 都是互联网中的一个网页，而互联网中的每一个网页都是通过网页中的 URL 链接扇出到另外的 URL 中的。这给爬虫带来的一个问题是，在抓取一个具体 URL 中的数据时，如何处理这个 URL 中扇出的 URL 链接？这就涉及网络爬虫数据的抓取策略了。

网络爬虫抓取策略是指在网络爬虫系统中决定 URL 在待抓取 URL 队列中排列顺序的方法。不同的网络抓取策略将对应不同的网页抓取过程，相应的抓取效率也有所不同，常见的网络爬虫抓取策略有如下 4 种，如表 5-4 所示。

表 5-4 不同抓取策略的特点比较

抓取策略	描述	特点
深度优先策略	从 URL 池中选择某 URL，然后按深度优先的思想遍历以该 URL 为根节点的所有 URL 的网页内容，然后取出 URL 池中的下一个 URL，继续上述策略循环至 URL 池遍历完	抓取深度大，但容易导致无限抓取，使得抓取过程无法收敛
广度优先策略	按照广度优先的搜索思想，逐层抓取 URL 池中的每一个 URL 的内容并将每一层的扇出 URL 纳入 URL 池中，按照广度优先的策略继续遍历	抓取宽度广，抓取过程容易控制，能有效地减轻服务器的负载，但容易造成 URL 大量聚集而导致 URL 池溢出
局部 PageRank 策略	借鉴 PageRank 的思想，在 URL 池和已抓取网页组成的网页集合中计算 URL 池中 PageRank 的值并以此进行排序，然后按照此排序顺序遍历各个 URL	网络环境中，由于广告链接、作弊链接的存在，易导致 PageRank 的值不能完全刻画其重要程度，从而导致实际抓取的数据无效
OPIC 策略	OPIC 策略将每个网页赋予相同的“金币”，每当下载某个页面 P，则将 P 拥有的“金币”平均分配给网页中包含的链接页面。待爬队列中的链接依“金币”排序	OPIC 计算速度快于局部 PageRank 策略，是一种较好的重要性衡量策略，适合实时计算场合

需要补充说明的是：

- 1) PageRank 是一种著名的链接分析算法，其功能是通过计算网页被其他网页链接指向的数量来表示其重要性，借此实现对每个网页的重要度排序。
- 2) OPIC 是“Online Page Importance Computation”的缩写，意为“在线页面重要性计算”。
- 3) 在实际操作中，用户设定 URL 的下意识就是收集与当前 URL 直接链接的有限扇出层级，甚至是该 URL 指定网页中的某个频道（模块）的扇出 URL，因此在实际的网络抓取过程中，除了配置必要的抓取策略以外，往往需要对每一个 URL 配置一个专门的抓取策略。基于这样的思路，网络爬虫的实际执行步骤是：①从 URL 池中选择某个 URL。②读取与该 URL 对应的专有抓取策略，按照此专有策略抓取该 URL 的内容。③依照事先约定的抓取策略（深度

优先、广度优先、局部 PageRank、OPIC 等) 选择下一个 URL。

网络爬虫的目的在于抓取指定网页的信息, 因此针对网络爬虫的评估一般需要从网络爬虫程序的开发者和抓取内容两个角度对网络爬虫进行评价, 具体如表 5-5 所示。

表 5-5 网络爬虫评价

评价视角	评价维度	评价细节
程序开发	高效性	一般的衡量标准为每秒钟抓取的网页数量, 每秒钟抓取的网页数据量越大, 爬虫程序越高效
	可扩展性	不同的网页具有不同的(模板)结构, 针对不同的应用场景, 网络爬虫数据抓取的需要也不一样, 均须网络爬虫具有良好的扩展性
	健壮性	网络爬虫程序必须具有良好的容错性, 能够正确处理相关的异常情况, 保障抓取过程正常进行
抓取内容	友好性	一方面, 网络爬虫程序应该易于管理 URL 池; 另一方面, 网络爬虫程序在抓取过程中应减少被抓取网站的网络负载
	抓取网页覆盖率	网络爬虫应具有较大的抓取网页覆盖率(指抓取的网页占整个互联网的比例), 抓取网页的覆盖率越大, 则表明抓取的网络大数据就可能越全面
	抓取网页及时性	应及时获取最新的网络数据, 以保持抓取的网络大数据的“活性”
	抓取网页重要性	应该抓取具有重要价值的网页, 使抓取过程具有较高的性价比

基于上述介绍, 可以看到, 对于一个具体的网络爬虫(软件)而言, URL 池中 URL 的数量及这些 URL 中数据更新的频率会直接影响到网络爬虫的计算复杂度和网络数据抓取效率。因此, 在大数据应用场景下, 使用分布式计算技术, 将网络数据抓取并行化, 已经成为一个必然的趋势。

所谓分布式网络数据抓取, 就是通过多个单机爬虫系统的有效协作和配合, 实现互联网大数据的数据抓取。显然, 分布式网络数据抓取至少涉及网络爬虫本身和(分布式)任务分配(策略、方法和技术), 而后者是完全独立于网络数据抓取这个技术领域的。

所谓任务分配就是探讨如何将一个给定的任务(URL 池中的数据抓取)分配给一个或多个合适的 Agent(单机的爬虫系统), 从而充分利用系统的资源, 提高系统的运作绩效。从系统优化的角度出发, 任务分配的问题是在满足资源约束的条件下, 根据给定的资源、任务及相应的性能评价, 最大化完成任务的问题。

从 Agent 间是否显式合作完成任务的角度而言, 任务分配可分为涌现式分配和约定式分配, 前者指的是当反应式 Agent 受到来源于环境或多 Agent 系统内部的刺激时, 会执行相应的任务, 从而达到任务分配的目的; 后者又可分为集中式任务分配和分布式任务分配。

所谓集中式任务分配指的是系统中存在一个 Controller, 其拥有系统的全局信息, 代表系统的整体利益来建立分配的最优方案, 之后将分配方案通知给系统中的各相关 Agent。集中式任务分配的思路一般有 4 种, 如表 5-6 所示。

表 5-6 集中式任务分配思路

研究视角	描述	缺陷
数学规划	清晰地揭示任务分配问题的目标和约束，容易被理解和分析，有利于系统特征较为明确的任务分配问题的建模和求解	计算量大且缺乏健壮性，任务的变化和系统中节点的增删，都需重建规划模型
智能优化	将任务分配为一个单目标或多目标寻优问题，该方法在实际应用中往往能够得到很好的效果	无理论保证，且容易陷入局部最优
搜索问题	将任务分配描述为给定约束条件下，按照一定的规则，在一定目标的驱动下寻找问题的最优解，该方法直观、易理解	全局搜索计算量大，甚至需要牺牲全局最优解来降低算法的复杂度
图论问题	利用图论方法建立任务和 Agent 间的匹配，从而产生有效的分配方案，该方法只需要用点和线就能直观地表达任务分配的问题	当任务和节点数都很大时，很难用清晰的图来刻画复杂任务分配问题的特征

在分布式任务分配中，不存在掌握全局信息的 Controller，Agent 的角色分配是随环境变化而动态变化的，任务分配由多个 Agent 共同参与、协商和竞争，或者各 Agent 之间根据对环境信息的感知，完全独立地进行任务的选择或调整。目前关于分布式任务分配的研究思路一般有 4 个方面，具体见表 5-7。

表 5-7 分布式任务分配思路

研究视角	描述	缺陷
基于行为激励	将环境和感知信息映射到 Agent 的智能行为模式，Agent 根据自身行为模式的改变，实现全局任务分配方案的自动调整	该方法适用于由自治性较强的 Agent 构成的多 Agent 任务分配问题，该类方法容错性好，但分配效率不高
基于市场机制	本质是多 Agent 系统中的 Agent 个体为求得利益最大值，在某种协议的基础上与其他 Agent 通过对话、协商来动态地分配任务	适用于显式通信模式下，Agent 之间通过协商来调整分配方案，该方法分配效率高，但系统通信负载较大
基于空闲链	基本思想是系统出现一个空闲资源时，需对该空缺重新进行分配，当该空缺被填充时，会导致系统中出现新的空缺，从而产生空闲链，带动整个系统自动实现动态再分配	重点解决 Agent 失效或通信中断等突发情况时系统应如何动态调整任务分配方案，该方法自适应性强，但使用时需要满足相应的条件
基于群智能	将任务分配问题建模为一个优化问题，然后利用群智能的寻优优势进行问题的求解	具有较好的灵活性和健壮性，分配效率高，可是仅适用于 Agent 行为简单但数量较多情况下的集中式或分布式任务分配问题

事实上，关于任务分配的现有研究，无论是集中式任务分配还是分布式任务分配，研究者普遍关注的重点是如何提高解的质量及求解效率，此处不再一一赘述。

5.4.3 开源网络爬虫

大数据时代网络数据的采集是进行大数据挖掘的前提，采集到的数据类型和质量对后续大数据价值的挖掘至关重要。网络大数据通常采用网络爬虫进行采集，开发网络爬虫采集网络大数据显得尤为重要。目前有非常多的开源网络爬虫可供开发人员使用，如 Nutch、Scrapy、

Larbin、Heritrix、JSpider、Crawler4j、WebSPHINX、Mecrator、PolyBot 等。本节将重点介绍几种具有代表性的开源网络爬虫，在实际开发中，可根据具体需求选择合适的爬虫框架。

(1) Nutch

Nutch 是一个用 Java 语言实现的开源搜索引擎，提供运行自己的搜索引擎所需的全文搜索和 Web 爬虫等全部工具。Nutch 支持分布式抓取，并有 Hadoop 支持，可以进行多机分布式抓取、存储和索引。Nutch 爬虫框架采用插件架构设计，具有高度模块化特性，且易扩展、伸缩性强，如通过插件可以实现对各种网页内容的解析，以及各种数据的采集、查询、集群、过滤等功能的扩展。Nutch 爬虫框架的很多模块都是采用配置文件的形式进行组织的，具有高度灵活性的同时也能保障整个框架具有极强的健壮性和可靠性。

关于 Nutch 的详细介绍请参见其官网 <http://nutch.apache.org/>。

(2) Scrapy

Scrapy 是用 Python 开发的一个快速的、高层次的 Web 抓取框架，用于抓取网页并从页面中提取结构化的数据，简单轻巧，并且非常方便。Scrapy 网络爬虫框架在信息提取时采用可读性更强的 xpath 来代替正则表达式，采用中间件的形式，方便编写统一的过滤器，并采用管道的方式实现抓取数据的持久化。它使用异步网络库来处理网络通信，结构清晰，并且包含了各种中间件接口，可以灵活地实现各种需求。使用者及开发者可以根据自己的实际需求自由地开发 spider 模块及各种中间件来完成对某类网站的定制化抓取，Scrapy 在数据的序列化方面也提供了多种实现，支持 JSON、CSV、XML 等。

关于 Scrapy 的详细介绍可参见其官网 <http://scrapy.org/>。

(3) Larbin

Larbin 是基于 C++ 的 Web 爬虫工具，拥有易于操作的界面，运行于 Linux 系统下。Larbin 能够跟踪页面的 URL 进行扩展抓取，为搜索引擎提供广泛的数据来源。利用 Larbin，我们可以轻易地获取和确定单个网站的所有链接，甚至可以镜像一个网站；也可以用它来建立 URL 列表群，如针对所有的网页进行 URL Retrive 后，进行 XML 链接的获取。Larbin 的特点是简单、可配置性强，一个基于 PC 的简单的 Larbin 爬虫每天可以获取 500 万个网页，非常高效。

关于 Larbin 的详细介绍可参见其官网 <http://larbin.sourceforge.net/>。

表 5-8 给出了上述 3 种开源爬虫的特点与区别，仅供参考。

表 5-8 3 种主流爬虫软件的特点与区别

爬虫	优点	缺点	适用场景	开发语言
Nutch	1) 集抓取、索引于一体。基于 Hadoop 的分布式系统 2) 存储层剥离，支持存储 HBase、Cassandra、MySQL 等数据库 3) 基于插件式设计，扩展和定制比较方便 4) 支持网页解析和索引，可以对接至 Solr，搭建通用的搜索引擎	Nutch 更侧重于索引，和 Hadoop 结合之后，会消耗更多的资源在非爬虫部分，抓取效率较低	不仅需要抓取数据，同时对索引也有一定的需求。大数量、考虑分布式的场景下可以直接使用 Nutch 的分布式解决方案	Java

(续)

爬虫	优点	缺点	适用场景	开发语言
Scrapy	1) 插件设计, 扩展性比较好 2) 爬虫规则定制简单 3) 支持抓取和抽取, 数据抽取结构化 4) 抽取支持 xpath 和 CSS 提取网页数据	1) 单机多线程实现, 默认不支持分布式, 分布式需要自己实现 2) 数据存储方案支持 Local filesystem、FTP、S3、Standard output, 默认无分布式存储解决方案 3) 默认中间过程网页不会保存, 只保存抽取结果	1) 没有分布式需求, 或者有其他分布式解决方案 2) 只需要抽取结果, 对原始网页不感兴趣	Python
Larbin	单纯的爬取功能, 简单, 单机效率高	1) 不支持分布式系统的抓取和存储 2) 功能相对简单, 提供的配置项也不够多 3) 不支持网页自动重访及更新功能	1) 只需要爬虫工具, 其他功能可通过其他方案解决 2) 硬件有限, 但对效率要求较高 3) 适合作为定制化爬虫系统的爬虫器部分	C++

5.4.4 应用提示

在大数据项目的建设过程中, 作为最关键的环节, 数据采集的数据源对象包括本单位的自营数据、其他单位的他营数据、互联网数据, 本节提到的外部数据是专门针对互联网数据而言的。网络爬虫是从互联网上抓取数据的有效手段。在具体操作过程中, 一般的流程是: 用户设定 URL 池和策略 (总体策略及每一个 URL 的策略), 爬虫软件系统 (单机的或分布式的) 依据设定的策略依次抓取 URL 池中的数据。在这个过程中, 有以下几个问题值得注意:

1) URL 的配置是一项需要用户、开发人员等共同进行的工作: 一方面需要从应用驱动的角度研判该 URL 数据是否有用; 另一方面需要从技术手段的角度评估该 URL 数据是否可以可信获取, 在此评估过程中会将新增爬虫的需求反馈给爬虫开发团队。

2) 网络爬虫软件往往是处于不断的迭代过程中的, 这一方面是因为 URL 网页的结构和编程方式方法没有统一的标准 (或者说虽有统一的编程语言标注, 但形式多样), 使得几乎没有可能用同一爬虫软件抓取所有类型网页的数据, 因此网络爬虫软件处于持续的迭代开发中。

3) 对于被抓取的网站而言, 出于对网络负载的考虑, 往往会设置一些策略和方法阻止爬虫系统的数据抓取活动, 这就意味着, 在爬虫策略的指定方面, 除了需要考虑从纯粹的技术流角度配置相应的策略以外, 还需要从抓取频度、网络代理等多个角度进行配置, 并且应当注意到爬虫技术和反爬虫技术从来都是一个此消彼长的关系。

4) 从某个具体的 URL 中获得的数据往往是一个很长的字符串, 而其中哪些部分用户会感兴趣需要有专门的技术手段进行分析。一般通过两种策略来进行: 一种是在网络爬虫抓取该 URL 数据的时候自动解析, 将感兴趣区域的数据保存下来, 而其他的不予保存; 另外一种是将该 URL 数据全文保存在数据库中, 后台程序另行自动读取、解析和转存。

5) 在实际工作中,往往需要对从 URL 中读取的数据,特别是感兴趣的数据,进行必要的预处理,为后续的分析提供高质量的数据基础。必不可少的预处理内容包括:去重、结构化、自动摘要、标签化等:

①去重的目的是将重复获取的内容剔除,仅保留其中的一个版本(这对于降低存储量非常有效)。进行去重操作的同时,一般会把重复次数、转载 URL 等信息记录下来。其中,重复次数可用于评估该 URL 数据的热度,转载的 URL 记录可用于评估该 URL 的原创或转载偏好。

②结构化的目的是将 URL 数据(往往是非结构化的文本)中的结构化信息提取出来,以便对此数据进行高层语义理解。这往往会涉及具体的目标应用,即根据目标应用场景及该 URL 数据的特点设计结构化的方式和方法。

③自动摘要的目的是将 URL 数据(往往比较长)以更短的文本方式加以描述,以便后续的用户进行分析回溯时能更扼要地了解每个 URL 的内容。

④标签化的目的是为该 URL 数据打上不同的标签,以便为后续进行数据分析和研判提供基础的数据画像。一般的标签包括底层语义标签(比如关键词标签,通过用户设定的热词,为该 URL 数据赋予不同的关键词属性)、情感语义标签(通过情感分析获得该 URL 数据的情感倾向)、高级语义标签(往往是与具体应用目标有关的一些高级语义,比如从 URL 数据中分析其中蕴含的风险信息,属于何种风险等)。不论是上述的哪一种标签,都是与应用相关的(耦合度不同而已),需要领域知识的支撑。

6) 在进行实时 URL 数据分析的时候,“热(新)词发现”和“主题发现”通常是比较常见的功能需求,前者专注于从实时的活性数据中发现新生的词汇;后者专注于从实时的活性数据中发现主题及各个主题下的热词,这对于事件的发现和分析具有重大的意义。

5.5 本章小结

美国管理学家、统计学家爱德华·戴明有一句名言是这样说的:除了上帝,任何人都必须用数据说话。事实上,从数据中发现知识、凝练智慧是人类文明进程中亘古不变的活动轨迹。

从数据中发现知识和洞见的一个基础条件是必须有数据。在大数据应用场景下,用于分析的数据散布于不同的利益主体的数据中心里,鉴于没有任何一家单位能够采集全人类的所有数据,甚至可以说,任何一个集团(哪怕是一家“航母”级企业)都没有拥有针对某个应用场景的所有数据,因此数据的获取是摆在大数据应用中每一个利益集团面前的第一个命题。

对于任何一个开展大数据应用的利益主体而言,数据的来源无外乎有两种:①内部自营数据。这类数据源于利益主体自己的平台,包括虚拟网络平台生产的数据,或者从利益主体自营的终端产生的数据。②外部数据。这里数据包括互联网数据和存储在其他利益集团,但在互联网上无法获得的其他利益主体的内部数据。因此,单纯就技术层面而言,内部自营数据可以通过 ETL 进行汇聚;互联网数据可以通过网络爬虫来获得;而其他利益主体的数据就

必须通过商务上的约定,然后通过 ETL 来获得,而显然这需要商务交换,或者在某一个商业模式的共享获益支撑下才可以完成。

在实际操作层面,数据的采集与整合远没有技术流说明那样的纯粹,原因如下:

1) 以内部自营数据为例,对于任何一个公司,特别是大型公司,如中国移动这样的航母级运营商,所有的数据均归属于不同的部门。从数据采集的角度而言,自然是希望能够采集不同部门的数据,但是各个部门从涉密、安全及部门的局部利益出发,未必有意愿和能力把自己所掌握的数据共享出来。另一方面,在为具体的客户单位开发大数据项目时,许多需要采集的数据是存放在不同的业务系统中的,这些业务系统的数据是否能够被获取,也是一个实际问题。还有一个普适的麻烦在于:许多单位的遗留数据是以非数字化的手段存放的,如何将既有的数据数字化、格式化也是开展大数据应用过程中经常会遇到的问题。

2) 互联网数据的获取也并不容易。任何一个网站出于负载均衡及相关利益的目的,并不是特别欢迎机器爬虫前往它的网站自动获取数据,因而会设置类似校验码这样的障碍。这对于单纯用于获得数据驱动的爬虫而言,显然是一个难以解决的技术问题。一种解决方案或许是通过一种商务合作的方式获得数据的交换。以新浪微博为例,虽然利用爬虫并基于一定的策略可以获得一些数据,但是如果期望获得的数据更全面、更高效,与新浪微博进行商务合作或许是最为方便的解决方案。

3) 还应当看到,特别是随着移动互联网的迅猛发展,如何从移动终端获得用户的第一手数据,也是大数据采集过程中的一个重要难题。这方面的数据获取往往需要在一个良好的商业模式及很多资源的共同支撑下才能得以实现。比如很多公司开发了很多的 APP 并以免费的方式提供给用户,用户在享用免费服务的同时将数据免费地提交上来。显然这样的数据采集方法,需要 APP 足够好,具有极大的黏性等,这已经不是简单的数据采集问题了。

4) 如前所述,开展大数据应用项目的第一个问题是数据的采集与整合,但是数据源的获取渠道太多,多得以至于无法在既有条件下全部自行开展相关工作。目前市场上出现了专门的数据服务公司,他们利用自己的技术手段,或者垄断的渠道入口来采集相关的数据(一般是面向某一个垂直领域)。因此在进行大数据采集的时候,与这样的公司展开合作也是一个可取之道。

5) 数据源的遴选和定义也是大数据项目开展之前的一个重要内容。尽管大数据分析需要更多、更全的数据,但是“数据能满足其既定的用途,它才能具有质量”。因此从哪里及如何获得更有价值的、有质量的数据,是大数据项目开展过程中在数据采集工作时不可回避也必须引起注意的问题。显然,这两个问题是在对应用有了极具敏锐的论域分析以后才能够解决的。

《百喻经》,也称《百句譬喻经》,是古天竺僧伽斯那以梵文撰写,后由其弟子求那毗地在 492 年译作的汉文。《百喻经》称“百喻”,就是指有一百篇寓言故事,每篇都分为喻体和法义两个部分,前者讲故事,后者阐述故事所显示的教诫。《三重楼喻》是《百喻经》的第 10 个,描述的是脍炙人口的“空中楼阁”的故事,其大意是说有一个财主期望在不建一、二楼

的情况下,直接建造第三层楼,用于比喻虚幻的事物或脱离实际的空想。

与舶来品对“空中楼阁”的批判不同,中国文化对于“空中楼阁”(或者类似的概念“海市蜃楼”)具有更多的客观描述和人文情怀,表达的甚至是一种通透。比如《史记·天官书》中提及“海旁蜃气象楼台,广野气成宫阙然”;唐代诗人宋之问所作《游法华寺》中提及“空中结楼殿,意表出云霞”;白居易《长恨歌》中提及“忽闻海上有仙山,山在虚无缥缈间。楼阁玲珑五云起,其中绰约多仙子”……

这对于大数据应用的提示或许在于:大数据的价值值得用任何方式期待和膜拜,只要有基础,所有的价值期望都不是虚无缥缈的“空中楼阁”,而显然,数据是最坚实的基础。正所谓“得数据者得天下”,大数据价值的实现将从数据采集开始。

本章参考文献

- [1] Abiteboul S, Preda M, Cobena G. Adaptive On-line Page Importance Computation [C]. Proceedings of the 12th International Conference on World Wide Web. ACM, 2003: 280-290.
- [2] Boldi P, Codenotti B, Santini M, et al. Ubcrawler: A scalable Fully Distributed Web Crawler [J]. Software: Practice and Experience, 2004, 34(8): 711-726.
- [3] Gantz J, Reinsel D. 2011 Digital Universe Study: Extracting Value From Chaos [M]. Framingham: IDC Go-to-Market Services, 2011: 1-12.
- [4] Holmstrom B, Milgrom P. Multitask Principal-agent Analyses: Incentive Contracts, Asset Ownership, and Job Design [J]. Journal of Law, Economics, & Organization, 1991: 24-52.
- [5] Page L. The PageRank Citation Ranking: Bringing Order to the Web [C]. Stanford InfoLab. 1999: 1-14.
- [6] Shkapenyuk V, Suel T. Design and Implementation of A High-performance Distributed Web Crawler [C]. Data Engineering, Proceedings, 18th International Conference on. IEEE, 2002: 357-368.
- [7] Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work and Think [M]. Hodder & Stoughton. 2013.
- [8] 杜绍森. Informatica PowerCenter 权威指南 [M]. 北京: 电子工业出版社, 2015.
- [9] 官建文. 中国移动互联网发展报告 [M]. 上海: 社会科学文献出版社, 2012.
- [10] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述 [J]. 计算机科学, 2008, 35(2): 1-5.
- [11] 涂子沛. 数据之巅: 大数据革命, 历史、现实与未来 [M]. 北京: 中信出版社, 2014.
- [12] 王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望 [J]. 计算机学报, 2013, 36(6): 1125-1138.
- [13] 张俊林. 这就是搜索引擎: 核心技术解密 [M]. 北京: 电子工业出版社, 2012.
- [14] 张宁, 贾自艳, 史忠植. 数据仓库中 ETL 技术的研究 [J]. 计算机工程与应用, 2002, 38(24): 213-216.

数据存储与管理

在本章行文之初，南京大学软件学院的刘嘉副教授对本章的内容梳理给予了很多建设性的意见；本章的写作及润色，得到了南京大学计算机科学与技术系及智能信息处理研究组的刘勇、张弛、韩军华、冯瀚洋、谭龙海等几位同学的协助，在此表示深深的谢意。

6.1 引言

没有纸张的时候，古人利用石子记数、结绳计数、龟壳记事。

没有纸张的时候，古人也用竹简记事，留下文字。

后来有了纸张，人们可以更便捷地记下文字、图形。

有了纸张后，人们还发明了纸带机，成为计算机的标准输入输出设备。

再后来，有了磁带、软盘、硬盘、光盘、U 盘……

显然，上述的“数”“事”“文字”“图形”等都是一种数据。存储介质的不断发展无疑能让人们更容易记录下更多的数据，同时也进一步倒逼人们去思考：如何更有效地组织数据，以便更好地利用这些数据，比如简单的检索、复杂的分析或更具视觉特征的展示。

数据存储涉及存储介质和数据组织方法，如同两条彼此缠绕的纽带螺旋式地推动着人类对数据存取方式的改进，而很显然的一点是：全电子计算机的发明如同是一个催化剂，使得数据的存储与管理方式持续进步，并不断发酵。

计算机最初主要用于数学计算，应用程序中用到的数据都要由程序员规定好数据的存储结构和存取方式等。一组数据只能面向一个应用程序，而不能实现多个程序的共享。不同程序之间不能直接交换数据，数据没有任何独立性。同时，当时的计算机也只能处理数字，不能处理字母和符号，而字母和符号恰恰就是数据处理中的主要处理对象。此外，当时还没有发明数据处理所需要的大容量存储器。20 世纪 50 年代初有两大技术进展推进了计算机的数据

处理能力：①发明了字符发生器，使计算机具有了能显示、存储与处理字母及各种符号的能力；②成功地将高速磁带机用于计算机作为存储器。但是磁带只能顺序读写，速度也慢，不是理想的存储设备。1956年，IBM公司和Remington Rand公司先后实验成功磁盘存储器方案，推出了商用磁盘系统。磁盘不但转速快、容量大，还可以随机读写，为数据处理提供了更加理想的大容量高速存储设备。有了这些硬件的支持，计算机数据处理便日益发展起来。

初期的数据处理软件只有文件管理（File Management）这种形式，即数据文件和应用程序一一对应，这种处理方式会导致数据依赖严重，即编写程序依赖于具体的数据，给程序的编制和维护都造成很大的麻烦。后来出现的文件管理系统（File Management System, FMS）便成为应用程序和数据文件之间的接口，一个应用程序通过FMS可以和若干文件打交道，在一定程度上增加了数据处理的灵活性。但这种方式仍以分散的、相互独立的数据文件为基础，数据冗余、数据不一致性、处理效率低等问题仍然不可避免，这些缺点在较大规模的系统中尤为突出。为了克服文件系统分散管理的弱点，实现对数据的集中控制和统一管理，一种全新高效的数据处理系统——数据库技术应运而生。表6-1简单地罗列了从20世纪60年代起数据库技术兴起至今的一个简单发展脉络。

表6-1 数据库发展脉络

时间	标志性技术	描述
60年代起	导航式数据库	1) 支持三级模式（外模式、模式、内模式） 2) 保证数据库系统具有数据与程序的物理独立性和一定的逻辑独立性 3) 用存取路径来表示数据之间的联系 4) 有独立的数据定义语言 5) 导航式的数据操纵语言
70年代起	关系型数据库	关系模型采用二维表格的结构表达实体类型和实体之间的联系，既简单又高效
80年代起	事务处理技术	1) 将对数据库的操作建模为一个个独立的事务，然后定义一系列的语义规范，确保数据库操作的完整性、安全性、并发性及容错性 2) 事务处理技术虽然诞生于数据库研究，但对于分布式系统、C/S结构中的数据管理与通信、容错和高可靠性系统都具有重要的意义

导航式数据库（Navigational Database）的特点描述如表6-1所示。事实上，导航式数据库具体有两种（个）：

1) 一种是IBM公司在IBM 360系列上推出的基于层次模型的数据库管理系统IMS（Information Management System）。

2) 另一种是巴赫曼（Charles W. Bachman）在通用电气公司主持设计与实现的基于网状模型的数据库管理系统IDS（Integrated Data System）。

基于层次的数据模型是有根的定向有序树，基于网络的数据模型对应的是有向图，这两种数据库奠定了现代数据库发展的基础。

巴赫曼主持设计与开发的网状数据库管理系统IDS于1964年推出后，成为最受欢迎的数据库产品之一，而且它的设计思想和实现技术被后来的许多数据库产品所仿效。同时，巴赫

曼本人还积极推动与促成了数据库标准的制定,并在1971年推出了第一个正式报告,即著名的DBTG报告,DBTG所提出的基本概念具有普遍意义,不但国际上大多数网状数据库管理系统,如IDMS、PRIME DBMS、DMSI70、DMS II和DMS 1100等都遵循或基本遵循DBTG模型的标准,而且对后来产生和发展关系型数据库技术也有很重要的影响。鉴于巴赫曼在数据库方面的巨大贡献,在1973年其本人获得图灵奖,这也是数据库领域第一位图灵奖获得者,并被誉为“网状数据库之父”。

埃德加·弗兰克·科德(Edgar Frank Codd)被誉为“关系型数据库之父”,并因为其在关系型数据库方面的贡献,获得1981年的图灵奖。关系型数据库模型采用二维表格结构表达实体类型和实体间的联系,既简单又高效,使得关系型数据库迅速占领了市场,至今仍然占统治地位。

关系型数据库最早由数据库之父埃德加·弗兰克·科德于1970年提出,当时科德正在IBM位于加州圣何塞的研究中心工作。但遗憾的是,科德的理论公开之后并没有立即被IBM采纳,因为当时IBM已经对一个IMS系统进行了大量的投资。不过,科德的理论很快启发了其他的公司和企业家去考虑如何进一步发展这一理论,其中一位是拉里·埃利森(Lawrence Ellison),他于1997年主导研制了世界上第一个商用关系型数据库管理系统,这就是著名的Oracle。

各大公司在关系型数据库的基本理论已经成熟,并被应用在关系型数据库管理系统(RDBMS)的实现和产品开发中的时候,遇到了一系列的技术问题:主要是在数据库的规模愈来愈大,结构愈来愈复杂,又有愈来愈多的用户共享数据库的情况下,如何保障数据的完整性(Integrity)、安全性(Security)、并行性(Concurrency),以及一旦出现故障后,数据库如何从故障中恢复(Recovery)。这些问题如果不能得到圆满解决,无论哪个公司的数据库产品都无法进入实用阶段,最终都不能被用户所接受。

针对上述问题,吉姆·格雷提出了“事务处理技术(Transaction Processing Technique)”,其基本思路是:①把对数据库的操作划分为“事务”的一个个基本执行单位:一个事务中的操作要么全部被执行,要么全部都不执行,以保障数据的完整性、一致性;②用户在对数据库发出操作请求时,系统对有关的数据元素进行“加锁”,操作完成后“解锁”,以保持事务之间的“并发性”和“完整性”;③建立系统运行日志,记载各事务的始点、终点,以及在事务中被更新的前后状态,以便在系统出现故障时将数据库恢复到系统故障前的正确状态,同时又能保留最后一次备份以来对数据库所做的所有修改;④鉴于一个事务可能同时涉及两个不同的数据库系统(特别是在分布式系统中),所以对数据库的任何更新都分为两个阶段提交。

因为格雷在事务处理技术上的创造性思维和开拓性工作,使他成为该技术领域公认的权威。因为格雷在数据库方面的巨大贡献,其在1998年获得图灵奖。这是数据库领域第三位图灵奖获得者,并被誉为“事务处理之父”。

大数据时代的来临,使得传统数据库尤其是关系型数据库无法很好地适应大数据带来的各种挑战,于是可以看到各大传统数据库航母公司在产品改进方面做了很多自适应的调整和完善。同时一个近年来极其流行的名字“NoSQL”跃然出世,“分布式”、“NoSQL”几乎成为大数据时代的一个技术标准被接纳和认可。

本章尝试从数据存储与管理的角度梳理数据库的发展脉络,以及为解决大数据引发的存储难题而应该具备的一些策略及技术选型思路,本章下面的结构安排如下:6.2节从集中与分布、SQL与NoSQL两个角度,对数据的组织与管理模式的相关思路和进展进行简单介绍,并简单地介绍了分布式文件系统的典型类别和典型NoSQL数据库,借此为数据库的技术选型提供借鉴;6.3节介绍数据存储及典型的数据存储系统的思路、关键技术及优势分析;6.4节介绍大数据环境下的一种存储服务模式,即云存储,这是对各方均有获益的商业部署实施模式,该节详细介绍云存储的思路、分类和优势等;6.5节对本章进行小结。

6.2 数据组织

如前节所述,数据的组织与管理经历了从最早期的人工管理阶段到文件管理阶段再到数据库管理阶段,显然,在这段历程中多年来都是数据库尤其是关系型数据库大行其道。不过大数据时代的来临,原先为“突破文件系统分散管理”而作为一种创举提出的数据库(技术)在处理“海量的、分布的、分散的数据”时却显得捉襟见肘。可以看到,虽然各大传统(关系型)数据库包括SQL Server、Oracle都给出了响应此需求的解决方案,但很显然:以开源驱动的分布式、小文件为核心的NoSQL向集中式、封闭性的关系型数据库发起了猛烈的挑战,而又由于其对大数据环境的天然适应性,传统的数据库市场,尤其是关系型数据库市场正在被其不断地蚕食。“去IOE”恰是在这个时期被热议、发酵的创举,这个创举是由中国的阿里巴巴提出并倡议的。

从2010年起,阿里巴巴在自己的IT架构中,去掉IBM的小型机、Oracle数据库、EMC存储设备,代之以自己在开源软件基础上开发的系统。这或许本来仅是一个企业家依据天赋异禀的“市场嗅觉”进行的决策或技术层面的革新思路。不过自2013年棱镜门事件之后,我国政府已经意识到政府数据安全性的重要性,也加强了政府数据安全方面的工作,“去IOE”与设备采购国产化、自主研发等口号挂钩,不可避免地带有了一定的政治色彩。政治性“去IOE”的提出,使得技术层次“去IOE”的争议和动作持续发酵。

“去IOE”的本质是用“分布式+开源”的架构替代“集中式+封闭”的架构,变成彻底的云计算服务模式。相对而言,“IE”的替代技术产品成熟,因此去“IE”要容易一些。而去“O”则困难得多,因为“O”与应用的耦合太过密切,且替代技术都不是特别成熟(相比较于去“IE”),因此如果真的要“去O”的话,理性的思路应当是:不盲动地政治性去“O”的同时,在技术性去“O”方面慎言、慎行。

6.2.1 集中与分布

信息资源在空间上集中配置的系统称为集中式系统，其主要优势体现在：①信息资源集中，管理方便，规范统一。②专业人员集中使用，有利于发挥他们的作用，便于组织人员的培训和提高工作效率。③信息资源利用率高。④系统安全措施实施方便。

典型的不足之处体现在：①随着系统规模的扩大和功能的提高，集中式系统的复杂性迅速增长，给管理、维护带来困难。②对组织变革和技术发展的适应性差，应变能力弱。③不利于发挥用户在系统开发、维护、管理方面的积极性与主动性。④系统比较脆弱，主机出现故障时可能使整个系统停止工作。

分布式是指利用计算机网络把分布在不同地点的计算机硬件、软件、数据等信息资源联系在一起，共同服务于同一个目标，实现相互通信和资源共享。分布式文件系统被认为是大数据的基础实施。所谓分布式文件系统（Distributed File System, DFS）是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过网络连接让多机器上的多用户分享文件和存储空间，因此 DFS 有时被称为网络文件系统（Network File System）。DFS 为文件系统提供了单个访问点和一个逻辑树结构，用户访问文件时并不直接访问底层数据存储块，而是以特定的网络协议与服务器进行沟通。

分布式文件系统的特点包括透明性和容错性。透明是指用户访问文件时并不知道文件的实际物理地址，或者说根本不知道文件系统是分布式存储的，在程序和用户看来就像在使用本地目录一样。容错是指即使系统中有一小部分的节点脱机，整体来说系统仍然可以持续运作而不会有数据损失。

当前比较流行的分布式文件系统包括：GFS、HDFS、Lustre、MogileFS、FastDFS、TFS、MooseFS、GridFS 等。它们各自适用于不同的领域，都不是系统级的分布式文件系统，而是应用级的分布式文件存储服务。以下将对上述系统进行简单介绍。

(1) GFS

Google 文件系统（Google File System, GFS）是 Google 公司为了能够存储以百亿计的海量网页信息而专门开发的文件系统。系统集群由一个 Master 节点和大量的 Chunk Server 节点构成，并被许多客户端（Client）访问。GFS 把文件分成 64MB 的块，缩小了元数据的大小，使 Master 节点能够非常方便地将元数据放置在内存中以提升访问效率。数据块分布在集群的机器上，使用 Linux 的文件系统存放，同时每块文件至少有 3 份以上的冗余。考虑到文件很少被删减或覆盖，文件操作以添加为主，充分考虑了硬盘线性吞吐量大和随机读点慢的特点。中心是一个 Master 节点，根据文件索引找寻文件块。系统保证每个 Master 都会有相应的复制品，以便于在 Master 节点出现问题时进行切换。在 Chunk 层，GFS 将节点失败视为常态，能够非常好地处理 Chunk 节点失效的问题。对于那些稍旧的文件，可以通过对它进行压缩，来节省硬盘空间，且压缩率惊人，有时甚至可以接近 90%。为了保证大规模数据的高速并行处理，GFS 引入了 MapReduce 编程模型，同时，由于 MapReduce 将很多烦琐的细节隐藏了起来，因

此也极大地简化了程序员的开发工作。

GFS 对大文件很友好, 吞吐量大, 但是延迟较高, 而且单一主控服务器限制了 GFS 可管理的文件数量。所以 Colossus 作为下一代 GFS 进行了相应的改进: 支持分布式主控服务器集群提升水平扩展性并支撑更多文件; 通过采用 Reed-Solomon 纠错码算法减少了 Chunk 数据的备份数量, 以此降低延迟。

(2) HDFS

HDFS(Hadoop Distributed File System) 最初是 Yahoo 模仿 Google 的 GFS 开发的分布式文件系统, 现在是 Apache 公司的 Hadoop 项目的核心子项目, 项目网站地址为 <http://hadoop.apache.org/hdfs/>, 相较于其他分布式文件系统, HDFS 具有高容错性的特点, 可以运行在廉价低配的服务器上; 而且 HDFS 可以为应用数据提供高吞吐量的访问, 因而适合开发超大规模数据集的服务和应用。此外, HDFS 通过降低可移植操作系统接口 (posix) 的要求达到流数据模式访问的目的。随着 Hadoop 的日渐流行, HDFS 目前也在各个应用场合被广泛使用。

HDFS 是一个主/从 (Master/Slave) 体系结构, 从最终用户的角度来看, 它就像传统的文件系统一样, 可以通过目录路径对文件执行 CRUD (Create、Read、Update 和 Delete) 操作。但由于分布式存储的性质, HDFS 集群拥有一个 NameNode 和多个 DataNode。NameNode 用于管理文件系统的命名空间和用户对数据访问相关信息的 master 服务节点, DataNode 用于管理和存储一部分实际的数据。客户端通过同 NameNode 和 DataNode 的交互访问文件系统。一个文件被分割成多个 64MB 大小的 block 数据块存储在部分 DataNode 中, 访问文件时客户端通过 NameNode 打开和关闭文件、获取文件的元数据, 而真正的文件 I/O 操作是直接和 DataNode 进行交互的。

HDFS 是用高移植性的 Java 语言编写的, 可以部署在大部分廉价的机器上, 因此专门用一台机器作为 NameNode 进行调度管理、其他机器作为 DataNode 存储和运算数据即可完成灵活简单的集群部署。

(3) Lustre(www.lustre.org)

Lustre 是一种并行分布式文件系统, 主要被设计用于大规模的集群计算。Lustre 这个名字由 Linux 和 Cluster 两个词混拼而成, 它是由 SUN 公司开发和维护的。该项目主要的目的是开发下一代的集群文件系统, 可以支持超过 10 000 个节点及 PB 级的数据存储。

Lustre 是开放源代码的集群文件系统, 采用 GPL 许可协议, 因此在超级计算机中得到广泛应用。世界 Top10 的超级计算机中, 有超过一半使用 Lustre 系统, 包括排名第二的 Titan 和 Sequoia。Lustre 文件系统具有高可扩展性, 可以支持数万个客户端节点、PB 级的存储及每秒 TB 级吞吐量的多机集群。因此 Lustre 成为包括气象地理、油气工业和生命科学等大规模数据应用的极佳选择。

(4) MogileFS(www.danga.com)

MogileFS 是一个开源的分布式文件系统, 用于组建分布式文件集群, 由 danga 团队开发。MogileFS 的存储引擎对应用完全透明, 同时它的每一个节点还可以作为一个轻量级

HTTP Server, 支持 GET 直接访问文件并支持接入时的负载均衡。MogileFS 由以下 3 个部分组成:

1) server 端: 包括 mogilefsd 和 mogstored 两个程序, 前者即 mogilefsd 的 tracker, 它将一些全局信息保存在数据库里; 后者即存储节点 (store node), 它其实是个 HTTP Daemon, 默认侦听在 7500 端口, 接受客户端的文件备份请求。在安装完后, 运行 mogadm 工具将所有的 store node 注册到 mogilefsd 的数据库里, mogilefsd 会对这些节点进行管理和监控。

2) utils (工具集): 主要是 MogileFS 的一些管理工具, 例如 mogadm 等。

3) 客户端 API: 目前只有 Perl API (MogileFS.pm)、PHP, 用这个模块可以编写客户端程序, 实现文件的备份管理功能。

(5) FastDFS (code.google.com/p/fastdfs)

FastDFS 是一款类似 Google FS 的开源分布式文件系统, 是纯 C 语言开发的, 它对文件进行管理, 功能包括: 文件存储、文件同步、文件访问 (文件上传、文件下载) 等, 解决了大容量存储和负载均衡的问题。特别适合于以文件为载体的在线服务, 如相册网站、视频网站等。FastDFS 服务器端有两个角色: 跟踪器 (Tracker) 和存储节点 (Storage), 前者主要做调度的工作, 在访问上起到负载均衡的作用; 后者存储节点存储文件, 完成文件管理的所有功能。

(6) TFS

TFS (Taobao File System) 是一个可扩展、高可用、高性能、面向互联网服务的分布式文件系统, 主要针对海量的非结构化数据, 它构筑在普通的 Linux 机器集群上, 可为外部提供高可靠和高并发的存储访问。TFS 为淘宝提供了海量的小文件存储, 通常文件大小不超过 1MB, 满足了淘宝对小文件存储的需求, 被广泛地应用于淘宝的各项应用中。它采用了 HA 架构和平滑扩容, 保证了整个文件系统的可用性和扩展性。同时扁平化的数据组织结构, 可将文件名映射到文件的物理地址, 简化了文件的访问流程, 一定程度上为 TFS 提供了良好的读写性能。

(7) MooseFS (derf.homelinux.org)

MooseFS (Moose File System) 是一个具备容错功能的网络分布式文件系统, 它将数据分布在网络中的不同的服务器上。MooseFS 通过 FUSE 使之看起来就像是一个 UNIX 的文件系统。但它还是不能解决单点故障的问题。开发语言是 Perl, 可跨平台操作。

(8) GridFS 文件系统

GridFS 是知名 NoSQL 数据库 MongoDB 的一个内置功能, 它提供了一组文件操作的 API 以利用 MongoDB 存储文件, GridFS 的基本原理是将文件保存在两个 Collection 中, 一个保存文件索引, 一个保存文件内容, 文件内容按一定的大小分成若干块, 每一块存在一个 Document 中, 这种方法不仅提供了文件存储, 还提供了与文件相关的一些附加属性 (比如 MD5 值、文件名等) 的存储。文件在 GridFS 中会按 4MB 为一个存储单位进行分块存储。

6.2.2 SQL 与 NoSQL

SQL 是 20 世纪 70 年代创建的一种基于关系型数据库管理系统 (Relational Database Man-

agement System, RDBMS) 模型的数据查询、操作语言,也是美国国家标准学会(ANSI)制定的标准。不同厂家的数据库产品在兼容此标准的同时一般会根据自己产品的特点对 SQL 进行一些改进和增强,于是就有了 SQL Server 的 Transact-SQL、Oracle 的 PL/SQL 等语言。

1970 年,在 IBM 工作的科德博士发表了里程碑性的论文《大型共享数据库数据的关系模型》,开创了关系型数据库的历史,不过出于多种原因,IBM 并没有立即采纳。1973 年,IBM 在外部的竞争压力下,开始加强在关系型数据库方面的投入。IBM 的丹·钱伯林博士被调到 San Jose 研究中心,加入新成立的项目 System R(基于科德提出的关系型数据库管理系统模型)。System R 项目包括研究高层的关系型数据系统(Relational Data System, RDS)和研究底层的存储系统(Research Storage System, RSS)两个小组,钱伯林担任 RDS 组的经理。RDS 实际上就是一个数据库语言编译器,由于科德提出的关系代数和关系演算过于数学化,影响了易用性。于是钱伯林选择了自然语言作为研究方向,其结果就是诞生了结构化英语查询语言(Structured English Query Language, SEQUEL),也就是现在脍炙人口的 SQL。因为钱伯林在 SQL 上的杰出贡献,其本人也被誉为“SQL 之父”。

System R 项目极具开创性:

- 1) 第一次实现了结构化查询语言,并使之成为日后的标准语言。
- 2) 第一个证明了关系型数据库管理系统可以提供良好的事务处理性能。
- 3) 系统设计中的决策理念及一些基本算法选择对后来的关系型系统都产生了积极的影响。

SQL 包括两个最主要的部件,即数据操作语言(DML)和数据定义语言(DDL)。前者主要用于执行查询、更新、插入和删除的语法(SQL 主要的 DML 语句有 SELECT、UPDATE、DELETE、INSERT INTO 等)。后者主要用于创建和删除数据库或表格,也可以定义索引(键),规定表之间的链接,以及施加表间的约束。SQL 主要的 DDL 语句有:CREATE DATABASE、ALTER DATABASE、CREATE TABLE、ALTER TABLE、DROP TABLE、DROP DATABASE、CREATE INDEX、DROP INDEX 等。

主流的关系型数据库有:Oracle、DB2、Microsoft SQL Server、Microsoft Access、MySQL 等。众所周知,关系型数据库中的表存储的是格式化后的数据结构,每个元组字段的组成都是一样的,尽管并不是每个元组都需要所有的字段,但数据库仍会为每个元组分配所有的字段,这样的结构优势很明显:

- 1) 便于在事务处理时保持数据的一致性。
- 2) 数据更新的开销很小。
- 3) 便于表与表之间进行连接等操作,从而进行复杂的查询等。

随着互联网的不断发展,特别是大数据时代的来临,各种类型的应用层出不穷,各类应用对数据库的要求也变得越来越苛刻。比如 Web2.0 网站要根据用户的个性化信息来实时生成动态页面和提供动态信息,此种需求对数据库并发负载的要求极高,传统的关系型数据库针

对每秒上万次的 SQL 查询或许还可以勉强应付,但是应付每秒上万次的 SQL 写请求,对于硬盘 IO 而言,这点几乎是个灾难。另一方面,当一个应用系统的用户量和访问量与日俱增的时候,传统的数据库却没有办法像 Web Server 和 App Server 那样简单地通过添加更多的硬件和服务节点来扩展性能和提高负载能力,这对于很多需要提供 24 小时不间断服务的网站来说,对数据库系统进行升级和扩展是极其困难的。

归纳而言,大数据时代的应用对数据的组织和管理提出了更多的需求(包括但不限于):

- 1) 低延迟的读写响应,借此提升用户的满意度。
- 2) 支撑海量的数据和流量。
- 3) 支持大规模集群的管理,借此响应系统管理员进行分布式应用的需求。
- 4) 降低运营成本。

显然,传统的关系型数据库虽然具备了天然的优势,但是在响应上述需求方面却显得捉襟见肘,具体体现在:

1) 扩展困难:由于存在类似 Join 这样的多表查询机制,使得数据库在扩展方面变得很艰难。

2) 读写慢:由于关系型数据库的系统逻辑非常复杂,因此在数据规模达到一定的量级时非常容易发生死锁等并发问题,从而导致其读写速度严重下滑。

3) 成本高:企业级数据库的 License 价格惊人,并且会随着系统规模的扩大而不断上升。

4) 支撑容量有限:现有关系型数据库解决方案还无法支撑 Google 这样海量的数据存储。

业界为了满足上面提到的几个需求,推出了多款新类型的数据库,并且由于它们在设计上和传统的 SQL 数据库相比有很大的不同,所以被统称为 NoSQL 数据库。NoSQL 有时也被称作是 Not Only SQL 的缩写,其以“键-值”对存储数据:它的结构不固定,每一个元组可以有不一样的字段,各个元组可以根据需要增加或减少自己的键值对,这样就不会局限于固定的结构,可以减少一些时间和空间的开销。同时在顶层设计方面, NoSQL 非常关注对数据高并发地读写和对海量数据的存储等,与关系型数据库相比, NoSQL 在架构和数据模型方面做了“减法”的同时在扩展和并发等方面做了“加法”。这些顶层设计策略使得 NoSQL 能够有效地满足上述大数据时代对数据组织与管理的需求,具体体现在:

1) 易扩展性: NoSQL 数据库种类繁多,但是一个共同的特点是去掉了关系型数据库的关系型特性。数据之间再无关系,这样就非常利于扩展。

2) 大数据量,高性能:得益于 NoSQL 的无关系性, NoSQL 数据库都具有非常高的读写性能,尤其在大数据量下,同样表现优秀。比如:一般 MySQL 使用 Query Cache(大粒度),每次表更新 Cache 就失效,在交互频繁的应用场景下, Cache 性能就不高。而 NoSQL 的 Cache 是记录级的(细粒度),从这个层面上来说 NoSQL 的性能就高了很多。

3) 灵活的数据模型: NoSQL 无须事先为要存储的数据建立字段,随时可以存储自定义的数据格式。而在关系型数据库里,增删字段是一件非常麻烦的事情。如果是非常大数据量的表,增加字段简直就是一个噩梦。

4) 高可用: 在不太影响性能的情况下, NoSQL 就可以方便地实现高可用性的架构。比如 Cassandra、HBase 模型, 通过复制模型就能实现高可用性。

或许是“时势造英雄”的原因, NoSQL 很好地适应了大数据时代对数据组织与管理的需求, 但是也存在着很多的不足, 共性的不足体现在: ①不提供对 SQL 的支持, 使得用户学习和应用的迁移成本很高(毕竟 SQL 作为一种标准已经成为共识)。②支持的特性不够丰富, 比如大多数 NoSQL 数据库都不支持事务, 也不像 MS SQL Server 和 Oracle 那样能提供诸如 BI 和报表附加的功能。③现有的 NoSQL 产品大都还处于初创期, 和关系型数据库几十年的完善程度不可同日而语。更大的麻烦在于: 在实际应用中, 每个产品都会根据自己所遵从的数据模型和 CAP 理念而有所不同, 这会给 NoSQL 像关系型数据库一样大行其道带来很大的困难。

对于任何一个分布式数据系统而言, 其三要素是一致性(Consistency, 指的是任何一个读操作总是能读取到之前完成的写操作结果, 也就是在分布式环境中, 多点的数据是一致的)、可用性(Availability, 指的是每一个操作总是能够在确定的时间内返回, 也就是系统随时都是可用的)、分区容忍性(Partition tolerance, 指的是在出现网络分区, 比如断网的情况下, 分离的系统也能正常运行)。根据 CAP 原理, 在分布式系统中, 这三个要素最多只能同时实现两点, 不可能三者兼顾。

互联网庞大的数据量和极高的峰值访问压力使得以增加内存、CPU 等节点性能的垂直伸缩方案(Scale Up)走入了死胡同, 使用大量廉价的机器组建水平可扩展集群(Scale Out)成为绝大多数互联网公司的必然选择。廉价的机器失效是正常的, 大规模的集群中节点之间的网络临时中断也是常见的, 因此在衡量一致性、可用性和分区容忍性时, 往往倾向于先满足后两者, 再用其他方法满足最终的一致性。不同的 NoSQL 产品尊崇不同的 CAP 理念, 比如在衡量 CAP 时, BigTable 选择了 CA, 用 GFS 来弥补 P; Dynamo 选择了 AP, C 弱化为最终一致性(通过 Quorum 或 read-your-write 机制)等。

2014 年 11 月 11 日, 淘宝“怒斩”571 亿交易额, 双十一的交易峰值已经达到 285 万笔/分钟, 相比 2013 年双十一期间 79 万笔/分钟的交易峰值, 2014 年系统的支撑能力达到了 2013 年的 3 倍以上, 用户整体支付体验相比去年也顺畅了不少, 页面打开的速度更快了, 支付等待的时间更短了。

但是凌晨 1 点, 多家电商平台反馈称, 支付宝出现支付故障, 用户无法正常支付。特别是从 11 日零点到 11 日 24 点这段时间内, 用户是不能得到退款的, 当时便有人戏称马云为了突破去年的成交额而“禁止退款”。但事实上, 且不说马云是不是需要通过这种方式来实现成交额的激增, 单看淘宝在系统架构上的设计, 在如此大规模并发的请求下, 退款将是一件非常困难的事情。

分布式架构有效地解决了海量数据高并发、高增长的问题, 对于淘宝这种以简单事务为主的 C2C 的业务模式来说也是最适合的。与此同时, 淘宝的系统架构也秉承了扩展性高于一

切、系统可用性高于一致性与适当放宽一致性约束等的原则。对于绝大多数应用场景来说，这种分布式系统是合适的、高效的。如果不是出现“双十一”这种前无古人、全球罕见的大规模交易场景，分布式系统堪称 C2C 业务的完美形态。

淘宝在系统架构中选择了扩展性与可用性，放弃了一致性，这依然符合“CAP 定律”的观点，意识到这一点的淘宝也采用了基于 MySQL 的分布式架构，从而完美地解决了高并发用户访问的难题。有道是“术业有专攻”，没有哪种架构是万能的，分布式也不是万能的。双十一是一面照妖镜，让我们看到分布式系统的强大，同时也看到集中式系统的稳健。即使你有勇气决定进行分布式的改造，但鉴于面临的风险、技术门槛、后期的运维等一系列问题，估计也只有像阿里这样的企业才能创造这样的神话。

目前主流的 NoSQL 产品有 MangoDB、MemBase、Hypertable、Apache Cassandra、BigTable、CouchDB、DynamoDB、SimpleDB 等，目前开源的 NoSQL 数据库有 Redis、Tokyo Cabinet/Tyrant、Apache Cassandra、Voldemort、MongoDB、Dynomite、HBase、CouchDB、Hypertable、Riak、Flare、Lightcloud、Scalaris、ThruDB 等，这些 NoSQL 数据库大致可以分为 3 类，具体请参见表 6-2。

表 6-2 主流 NoSQL 之间的区别

NoSQL 数据库分类	目标（特点）	代表性产品	
		非开源产品	开源产品
高性能 Key- Value 数据库	具有极高的并发读写性能	MemBase DynamoDB SimpleDB	Redis、Dynomite Tokyo Cabinet/Tyrant Flare、Riak Lightcloud、Scalaris
面向文档的非关系型数据库	虽没有很高的并发读写性能，但在保证海量数据存储的同时，具有良好的查询性能		MongoDB CouchDB ThruDB
面向 scale 能力的 NoSQL 数据库	以提供高扩展性和可用性为目的，例如可以不停机地添加更多的数据节点、删除数据节点等	BigTable	Apache Cassandra Hypertable Voldemort HBase

以下简单介绍各类主流的 NoSQL 数据库：

(1) Redis

Redis 本质上是一个 Key-Value（键 - 值）类型的内存数据库（单个 Value 的最大限制是 1GB），整个数据库系统加载在内存当中进行操作，定期通过异步操作把数据库中的数据 flush 到硬盘上进行保存。因为是纯内存操作，Redis 的性能非常出色，每秒可以处理超过 10 万次读写的操作。除此以外，Redis 还支持保存 List 链表和 Set 集合的数据结构，并支持对 List 和 Set 进行各种操作。因此 Redis 可以用来实现很多有用的功能，比方说用它的 List 来做 FIFO 双向链表，实现一个轻量级的高性能消息队列服务，用它的 Set 可以做高性能的 tag 系统等。Redis 的主要缺点是数据库的容量会受到物理内存的限制，不能用作海量数据的高性能读写，并且

它没有原生的可扩展机制,要依赖客户端来实现分布式读写,因此 Redis 适合的场景主要局限在较小数据量的高性能操作和运算上。

(2) MemBase

MemBase 是集群环境下的内存数据库,目前已更名为 CouchBase。MemBase 是由 NorthScale 和 Zynga 合作组织建立的项目, NorthScale 是广泛使用的缓存系统 MemCached 的制造商, Zynga 则是著名的社交游戏开发商。MemBase 起源于 Zynga 的实际需求:在社交游戏的环境下,需要高速、可靠且支持高吞吐量的存储系统,尤其是对写操作的效率要求很高。MemBase 就是在这种需求背景下产生的,其兼容 MemCached 协议,由 C、C++、Erlang 和 Python 混合语言写成。

MemBase 通过“虚拟桶”的方式对数据进行分片,其将所有数据的主键空间映射到 4096 个虚拟桶中,并在“虚拟桶映射表”中记载每个虚拟桶主数据及副本数据的机器地址, MemBase 对“虚拟桶映射表”的更改采用两阶段提交协议来保证其原子性。MemBase 中的所有服务器地位都是平等的,并不存在一个专门进行管理功能的 Master 服务器,但是其数据副本的管理采用了 Master-Slave 的模式。每个虚拟桶都有一台服务器作为主数据存储地,这台服务器负责响应客户端请求,副本存放在其他服务器内存中,其副本个数可以通过配置来指定。

MemBase 作为内存数据库,在架构设计上有比较完善的系统可用性保障措施,但是这种方法的缺点是所有的副本数据都存放在内存中,所以存储成本较高。

(3) Tokyo Cabinet 和 Tokyo Tyrant

TC 和 TT 的开发者是日本人平林干雄 (Mikio Hirabayashi),主要应用在日本最大的 SNS 网站 mixi.jp 上,前者是一个高性能的存储引擎,而后者提供了多线程高并发服务,每秒可以处理 4 到 5 万次读写操作。TC 除了支持 Key-Value 存储之外,还支持类似关系型数据库表结构的 Hash table 数据类型并支持基于 column 的条件查询、分页查询和排序功能,基本上相当于支持单表的基础查询功能了,这也是 TC 受到大家欢迎的主要原因之一。TC 主要的缺点是没有 scale 的能力,如果单机无法满足要求,只能通过主从复制的方式扩展,另外有人提到 TC 的性能会随着数据量的增加而下降,当数据量达到上亿条以后,性能会有比较明显的下降。

(4) Flare

Flare 是日本第二大 SNS 网站 green.jp 开发的,简而言之,Flare 就是给 TC 添加了 scale 功能,同时给 TC 另外写了网络服务器(替换掉 TT),Flare 在网络服务器端之前添加了一个 node server 用于管理后端的多个服务器节点,因此可以动态地添加数据库服务节点、删除服务器节点等。Flare 唯一的缺点就是它只支持 memcached 协议,因此当使用 Flare 的时候,就不能使用 TC 的 table 数据结构了,只能使用 TC 的 Key-Value 数据结构存储。

(5) DynamoDB

DynamoDB 是亚马逊 2012 年初发布的 Key-Value 模式的 NoSQL 存储平台,最初是为了满足公司内部需求的一个私有的分布式存储系统,产品目标是提供一个共享型的数据库云服务。DynamoDB 建立在 SSD 结构上,并能自动实现数据添加功能。具体做法是:将所有主键的散列

数值空间组成一个首尾相接的环状序列，为每台机器随机赋予一个散列值，而该机器就负责存储落在一段散列空间内的数据。数据定位使用一致性散列：对于一个数据，首先计算其散列值，根据其所落在的某个区段，顺时针进行查找，找到第一台机器，该机器就负责存储数据，对应的存取操作及冗余备份等操作也由其负责，以此来实现数据在不同机器之间的动态分配。DynamoDB 的顶层设计天然没有单点，每个实例由一组节点组成，从应用的角度看，实例提供 IO 能力。一个实例上的节点可能位于不同的数据中心内，这样如果一个数据中心出现问题也不会导致数据丢失。同时还可根据应用类型优化可用性、容错性和高效性配置去中心化、人工管理工作少。但是其存在的明显缺陷是可扩展性较差。由于增加机器需要给机器分配 DHT (Distributed Hash Table) 算法所需的编号，操作复杂度较高，且每台机器都存储了整个集群的机器信息及数据文件的 Merkle Tree 信息，机器的最大规模只能到几千台。

(6) MongoDB

MongoDB 是一个介于关系型数据库和非关系型数据库之间的产品，也是非关系型数据库当中功能最丰富、最像关系型数据库的。MongoDB 主要解决的是海量数据的访问效率问题，根据官方的文档，当数据量达到 50GB 以上的时候，MongoDB 的数据库访问速度是 MySQL 的 10 倍以上。MongoDB 的并发读写效率并不是特别出色，根据官方提供的性能测试，大约每秒可以处理 0.5 万到 1.5 次读写请求。MongoDB 支持的数据结构非常松散，是类似 JSON 的 bson 格式，因此可以存储比较复杂的数据类型。MongoDB 最大的特点是它支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系型数据库单表查询的绝大部分功能，而且还支持对数据建立索引，因为这个原因，很多项目都考虑用 MongoDB 替代 MySQL 来实现不是特别复杂的 Web 应用。

(7) CouchDB

CouchDB 是一个开源的面向文档的数据库管理系统，可以通过 RESTful JavaScript Object Notation (JSON) API 访问。术语“Couch”是“Cluster Of Unreliable Commodity Hardware”的首字母缩写，它反映了 CouchDB 的设计目标，CouchDB 具有高度可伸缩性的特点，提供了高可用性和高可靠性，即使运行在容易出现故障的硬件上也是如此。

与现在流行的关系型数据库服务器不同，CouchDB 是围绕一系列语义上自包含的文档而组织的。CouchDB 中的文档是没有模式的，也就是说并不要求文档具有某种特定的结构。CouchDB 的这种特性使得它相对于传统的关系型数据库而言，有自己的适用范围。一般来说，围绕文档来构建的应用都比较适合使用 CouchDB 作为其后台存储。CouchDB 强调其中所存储的文档，在语义上是自包含的。这种面向文档的设计思路，更贴近很多应用的真实情况。对于这类应用，使用 CouchDB 的文档来进行建模，会更加自然和简单。与此同时，CouchDB 也提供基于 MapReduce 编程模型的视图来对文档进行查询，可以提供类似于关系型数据库中 SQL 语句的能力。CouchDB 对于很多应用来说，提供了关系型数据库之外的更好的选择。

(8) Cassandra

Cassandra 是 Facebook 在 2008 年开发出来的，随后 Facebook 自己使用 Cassandra 的另外一

个不开源的分支,而开源的 Cassandra 主要由 Amazon 的 Dynamite 团队来维护,并且 Cassandra 被认为是 Dynamite2.0 版本。目前除了 Facebook 之外, Twitter 和 digg.com 都在使用 Cassandra。Cassandra 的主要特点就是它不是一个数据库,而是由一堆数据库节点共同构成的一个分布式网络服务,对 Cassandra 的一个写操作,会被复制到其他节点上去,对 Cassandra 的读操作,也会被路由到某个节点上面去读取。对于一个 Cassandra 群集来说,扩展性能是比较简单的事情,只需在群集里面添加节点就可以了。Cassandra 也支持比较丰富的数据结构和功能强大的查询语言,和 MongoDB 比较类似,查询功能比 MongoDB 稍弱。

Cassandra 的单个节点的并发读写性能不是特别好,有文章说评测下来 Cassandra 每秒只能处理大约不到 1 万次读写请求,不过单纯评价 Cassandra 单个节点的性能没有太大的意义,因为对于一个真实的分布式数据库访问系统而言,其并发性能取决于整个系统的节点数量、路由效率等综合因素,而不仅仅是单节点的并发负载能力。

(9) Voldemort

和 Cassandra 类似, Voldemort 也是面向解决 scale 问题的分布式数据库系统, Voldemort 来自于 LinkedIn 这个 SNS 网站。Voldemort 官方给出 Voldemort 的并发读写性能在每秒超过了 1.5 万次。

从 Facebook 开发 Cassandra、LinkedIn 开发的 Voldemort 可以大致看出国外大型 SNS 网站对于分布式数据库、特别是对数据库的 scale 能力方面的需求都极其迫切。前面提到过, Web 应用的架构当中, Web 层和 APP 层相对来说都很容易横向扩展,唯有数据库是单点的,极难扩展, Facebook 和 LinkedIn 在非关系型数据库的分布式方面探索了一条很好的道路,值得借鉴。

(10) BigTable

BigTable 是一种针对海量结构化或半结构化数据的存储模型,是一个稀疏的、分布式的、持久化存储的多维度排序 Map。BigTable 的设计目的是快速且可靠地处理 PB 级别的数据,并且能够部署在上千台机器上。本质上说, BigTable 是一个键值 (Key-Value) 映射。按作者的说法, BigTable 是一个稀疏的、分布式的、持久化的、多维的排序映射。BigTable 的键有三维,分别是行键 (Row Key)、列键 (Column Key) 和时间戳 (Timestamp),行键和列键都是字节串,时间戳是 64 位整型;而值是一个字节串。可以用 (Row: String, Column: String, Time: int64)→String 来表示一条键值对记录。

行键可以是任意字节串,通常有 10~100 字节。行的读写都是原子性的。BigTable 按照行键的字典序存储数据。BigTable 的表会根据行键自动划分为片 (tablet),片是负载均衡的单元。最初的表都只有一个片,但随着表的不断增大,片会自动分裂,片的大小控制在 100~200MB。行是表的第一级索引,我们可以把该行的列、时间和值看成一个整体,简化为一维键值映射。

列是第二级索引,每行拥有的列是不受限制的,可以随时增加或减少。为了方便管理,列被分为多个列族 (column family, 访问控制的单元),一个列族里的列一般存储相同类型的数据。一行的列族很少发生变化,但是列族里的列可以随意添加或删除。列键按照 family:qualifier 格式命名。我们可以将列拿出来,将时间和值看成一个整体,简化为二维键值映射。

时间戳是第三级索引。BigTable 允许保存数据的多个版本,版本区分的依据就是时间戳。

时间戳可以由 BigTable 赋值, 代表数据进入 BigTable 的准确时间, 也可以由客户端赋值。数据的不同版本按照时间戳降序存储, 因此先读到的是最新版本的数据。我们加入时间戳后, 就得到了 BigTable 的完整数据模型。

BigTable 使用集群管理系统来调度任务、管理资源、监测服务器状态并处理服务器故障。BigTable 使用 GFS 来存储数据文件和日志, 数据文件采用 SSTable 格式, 它提供了关键字到值的映射关系。BigTable 使用分布式的锁服务 Chubby 来保证集群中主服务器的唯一性、保存 BigTable 数据的引导区位置、发现 Tablet 服务器并处理 Tablet 服务器的失效、保存 BigTable 的数据模式信息、保存存取控制列表。Spanner 是 Google 的下一代 BigTable, 也是第一个全球级的分布式数据库, 可以将千亿规模的数据部署到世界范围内数百个数据中心的百万台服务器中, 利用 GPS 和原子钟实现全球规模数据的一致性和实时性。

BigTable 是建立在 Google 自有的 GFS 基础上的, 技术相对比较保密。但根据 Google 发表的论文, 业界已经有很多它的开源实现。其中比较成熟的有 HBase、HyperTable 等, 实现思路与 BigTable 类似。此处不再详细介绍。

其他 NoSQL 产品在此不再一一罗列, 下面仅给出所有产品的官网地址:

□ **SimpleDB**: <http://aws.amazon.com/cn/simplifiedb/>

□ **Hypertale**: <http://www.hypertable.org/3>

□ **Dynomite**: <https://github.com/moonpolysoft/dynomite/wiki/dynomite-framework>

□ **HBase**: <http://hbase.apache.org/>

□ **CouchDB**: <http://couchdb.apache.org/>

□ **Hypertable**: <http://www.hypertable.org/>

□ **Riak**: <http://basho.com/products/#riak>

□ **Lightcloud**: <http://opensource.plurk.com/LightCloud/>

□ **Scalaris**: <http://code.google.com/p/scalaris/>

□ **ThruDB**: <http://code.google.com/p/thrddb/>

6.3 数据存储

数据存储是数据在加工过程中产生的临时文件或加工过程中需要查找的信息。数据以某种格式记录在计算机内部或外部的存储介质上。数据存储要命名, 这种命名要反映出信息特征的组成含义。数据流是反映系统中流动的数据, 表现出动态数据的特征; 数据存储是反映系统中静止的数据, 表现出静态数据的特征。

常用的存储介质为磁盘和磁带。数据存储组织方式因存储介质而异。在磁带上的数据仅按顺序文件方式存取; 在磁盘上则可按使用要求采用顺序存取或直接存取的方式。数据存储方式与数据文件组织密切相关, 其关键在于建立记录的逻辑与物理顺序间的对应关系, 确定存储地址, 以提高数据存取速度。

早期比较成熟的网络存储结构大致分为3种：直连式存储（Direct Attached Storage, DAS）、网络连接式存储（Network Attached Storage, NAS）和存储网络（Storage Area Network, SAN），大致描述如下。

（1）直连式存储（DAS）

在直连式存储（DAS）中，主机与主机之间、主机与磁盘之间采用 SCSI 总线通道或 FC 通道、IDE 接口实现互联，将数据存储扩展到多台主机、多个磁盘。这种存储方式与我们普通的 PC 存储架构一样，外部存储设备都是直接挂接在服务器内部总线上的，数据存储设备是整个服务器结构的一部分。DAS 方式主要适用于以下环境：

□ 小型网络

因为网络规模较小，数据存储量小，且也不是很复杂，采用这种存储方式对服务器的影响不会很大。并且这种存储方式十分经济，适合拥有小型网络的企业用户。

□ 地理位置分散的网络

虽然企业总体网络规模较大，但在地理分布上很分散，通过 SAN 或 NAS 在它们之间进行互联非常困难，此时各分支机构服务器也可采用 DAS 方式，这样可以降低成本。

□ 特殊应用服务器

一些特殊应用服务器，如微软的集群服务器或某些数据库使用的原始分区，均要求存储设备直接连接到应用服务器。

在服务器与存储的各种连接方式中，DAS 曾被认为是一种低效率的结构，而且也不方便进行数据保护。直连式存储无法共享，因此经常出现的情况是某台服务器的存储空间不足，而其他一些服务器却有大量的存储空间处于闲置状态却无法利用。如果存储不能共享，也就谈不上容量分配与使用需求之间的平衡。

DAS 结构下的数据保护流程相对复杂，如果做网络备份，那么每台服务器都必须单独进行备份，而且所有的数据流都要通过网络传输。如果不做网络备份，那么就要为每台服务器都配备一套备份软件和磁带设备，所以说备份流程的复杂度会大大增加。

若要拥有高可用性的 DAS 存储，降低解决方案的成本是第一要务，如 LSI 的 12Gbit/s SAS，它有 DAS，通过 DAS 能够很好地为大型数据中心提供支持。对于大型的数据中心、云计算、存储和大数据，所有这一切都对 DAS 方式的存储性能提出了更高的要求。云和企业数据中心数据的爆炸性增长也推动了市场对于可支持更高速数据访问的高性能存储接口的需求，LSI 12Gbit/s SAS 正好能够满足这种性能增长的要求，它可以提供更高的 IOPS 和更强的吞吐能力，12Gbit/s SAS 提供了更高的写入的性能，并且提高了 RAID 的整个综合性能。

与直连式存储架构相比，共享式的存储架构，比如 SAN 或 NAS，都可以更好地解决上述问题。于是乎我们看到 DAS 被淘汰的进程越来越快了。可是到目前为止，DAS 仍然是服务器与存储连接的一种常用的模式。事实上，DAS 不但没有被淘汰，近几年似乎还有回潮的趋势。

（2）网络连接式存储（NAS）

NAS 的存储方式全面改进了以前低效的 DAS 的存储方式。它采用独立于服务器，单独为

网络数据存储而开发的一种文件服务器来连接存储设备,自形成一个网络。这样数据存储就不再是服务器的附属,而是作为独立的网络节点存在于网络之中了,可由所有的网络用户共享。NAS 的优点有如下几点:

- 1) 真正的即插即用: NAS 是独立存在于网络之中的存储节点,与用户的操作系统平台无关,是真正的即插即用。
- 2) 存储部署简单: NAS 不依赖于通用的操作系统,而是采用一个面向用户设计的,专门用于数据存储的简化操作系统,内置了与网络连接所需要的协议,因此使整个系统的管理和设置较为简单。
- 3) 管理容易且成本低: NAS 的数据存储方式是基于现有的企业 Ethernet 而设计的,按照 TCP/IP 协议进行通信,以文件的 I/O 方式进行数据的传输。

NAS 的缺点体现在存储性能较低及可靠度不高两个方面。

(3) 存储网络 (SAN)

1991 年,IBM 公司在 S/390 服务器中推出了 ESCON (Enterprise System Connection) 技术。它是基于光纤介质,最大传输速率达 17MB/s 的服务器访问存储器的一种连接方式。在此基础上,进一步推出了功能更强的 ESCON Director (FC Switch),构建了一套最原始的 SAN 系统。

SAN 的存储方式创造了存储的网络化。存储网络化顺应了计算机服务器体系结构网络化的趋势。SAN 的支撑技术是光纤通道 (FC Fibre Channel) 技术。它是 ANSI 为网络和通道 I/O 接口建立的一个标准集成。FC 技术支持 HIPPI、IPI、SCSI、IP、ATM 等多种高级协议,其最大的特性是将网络和设备的通信协议与传输物理介质隔离开,这样,多种协议即可在同一个物理连接上同时传送。SAN 的硬件基础设施是光纤通道,用光纤通道构建的 SAN 由以下 3 个部分组成。

- 1) 存储和备份设备: 包括磁带、磁盘和光盘库等。
- 2) 光纤通道网络连接部件: 包括主机总线适配卡、驱动程序、光缆、集线器、交换机、光纤通道和 SCSI 间的桥接器。
- 3) 应用和管理软件: 包括备份软件、存储资源管理软件和存储设备管理软件。

SAN 的优势体现在:

- 1) 网络部署容易。
- 2) 高速存储的性能。因为 SAN 采用了光纤通道技术,所以它具有更高的存储带宽,存储性能明显提高。SAN 的光纤通道使用全双工串行通信原理传输数据,传输速率高达 1062.5Mbit/s;
- 3) 良好的扩展能力。由于 SAN 采用了网络结构,扩展能力更强。光纤接口提供了 10km 的连接距离,这使得实现物理上的分离,不在本地机房的存储变得非常容易。

存储应用最大的特点是没有标准的体系结构,DAS、NAS 和 SAN 这三种存储方式共存,互相补充,已经很好地满足了企业的信息化应用。从连接方式上对比,DAS 采用了存储设备直接连接应用服务器的方式,具有一定的灵活性和限制性;NAS 通过网络技术连接存储设备和应用服务器,存储设备位置灵活,随着万兆网的出现,传输速率有了很大的提高;SAN 则是通过光纤通道 (Fibre Channel) 技术连接存储设备和应用服务器,具有很好的传输速率和扩

展性能。三种存储方式各有优势、相互共存, 占到了磁盘存储市场的 70% 以上。由于 SAN 和 NAS 产品的价格远高于 DAS, 因此许多对价格费用敏感的用户仍然选择低效率的直连式存储而不是高效率的共享式存储作为技术选型。

随着全球非结构化数据的快速增长, 针对结构化数据设计的这些传统存储结构在性能、可扩展性等方面都难以满足要求, 因此逐渐出现了集群存储、集群并行存储、P2P 存储、面向对象存储等多种存储结构。

1) 集群存储就是将若干个普通性能的存储系统联合起来组成“存储集群”。集群存储采用开放式的架构, 具有很高的扩展性, 一般包括存储节点、前端网络、后端网络三个构成元素, 每个元素都可以非常容易地进行扩展和升级而不用改变集群存储的架构。

2) 集群并行存储采用了分布式文件系统与并行文件系统相结合的方式, 容许客户端和存储直接打交道, 借此来提高并行或分区 I/O 的整体性能, 特别是读取操作密集型及大型文件的访问, 一般而言, 集群存储多用于大型数据中心或高性能计算中心。

3) P2P 存储用 P2P 的方式在广域网中构建大规模的分布式存储系统。从体系结构来看, 系统采用无中心结构, 节点之间对等, 通过互相合作的方式来完成任务。用户通过该平台自主寻找其他节点进行数据备份和存储空间交换, 从而构建大规模存储交换的系统平台。

4) 面向对象存储是 SAN 和 NAS 的有机结合。在面向对象存储中, 文件系统中的用户组件部分基本与传统文件系统相同, 但是文件系统中的存储组件部分被下移到智能存储设备上, 于是用户对于存储设备的访问接口由传统的块接口变为对象接口, 一般的认同是: 面向对象存储是存储系统的一个发展趋势。

6.4 云存储

云存储是一种网络在线存储 (Online Storage) 的模式, 即把数据存放在通常由第三方托管的多台虚拟服务器中, 而非专属的服务器上。托管 (Hosting) 公司营运大型的数据中心, 需要实施数据存储托管的公司, 通过向其购买或租赁存储空间的方式, 来满足数据存储的需求。数据中心营运商根据客户的需求, 在后端准备存储虚拟化的资源, 并将其以存储资源池 (Storage Pool) 的方式提供给客户, 客户便可自行使用此存储资源池来存放文件或对象。实际上, 这些资源可能被分布在众多的服务器主机上。云存储这项服务可通过 Web 服务应用程序接口 (API) 或 Web 化的用户界面来访问。

当我们使用某一个独立的存储设备时, 我们必须非常清楚这个存储设备的型号、接口和传输协议, 必须清楚地知道存储系统中有多少块磁盘, 分别是什么型号、多大容量, 必须清楚存储设备和服务器之间采用什么样的连接线缆。为了保证数据安全和业务的连续性, 我们还需要建立相应的数据备份系统和容灾系统。除此之外, 对存储设备进行定期的状态监控、维护、软硬件更新和升级也是必需的。如果采用云存储, 那么上面所提到的一切对使用者来

讲就都不需要了。云存储系统中的所有设备对使用者来讲都是完全透明的,任何地方的任何一个经过授权的使用者都可以通过一根接入线缆与云存储连接,对云存储进行数据访问。

云存储已经成为未来存储发展的一种趋势,采用云存储模式具有相当多的优势:

1) 节约成本。云存储从短期和长期来看,最大的特点就是可以为小企业节约成本。因为如果小企业想要在他们自己的服务器上存储,那就必须购买相应的硬件和软件,要知道它们是多么的昂贵。接着,企业还要聘请专业的IT人士,负责这些硬件和软件的维护工作,并且还要更新这些设备和软件。通过云存储,服务器商可以服务成千上万的中小企业,并可以划分不同的消费群体服务。它可以节省一个初创公司为拥有最新、最好的存储而花费的一部分成本,来帮助初创公司减少不必要的成本预算。相比传统的存储扩容,云存储架构采用的是并行扩容的方式,当客户需要增加容量时,可按照需求采购服务器,简单地增加即可实现容量的扩展:新设备仅需安装操作系统及云存储软件,打开电源接上网络,云存储系统便能自动识别,自动把容量加入存储池中完成扩展,扩容环节无任何限制。

2) 更好地备份本地数据并可以异地处理日常数据。如果你的办公场所发生自然灾害,由于你的数据是异地存储的,因此它是非常安全的。即使自然灾害让你不能通过网络访问到数据,但是数据依然存在。如果问题只出现在你的办公室或你所在的公司,那么你可以随便去一个地方用你的笔记本来访问重要数据和更新数据。它可以让你在恶劣的条件下依然保持工作。在以往的存储系统管理中,管理人员需要面对不同的存储设备,不同厂商的设备均有不同的管理界面,这就使得管理人员要了解每个存储的使用状况(容量、负载等),工作复杂而繁重。而且,传统的存储在硬盘或是存储服务器损坏时,可能会造成数据丢失,而云存储则不会,如果硬盘坏掉,数据会自动迁移到别的硬盘,大大减轻了管理人员的工作负担。对云存储来说,再多的存储服务器,在管理人员眼中也只是一台存储器,每台存储服务器的使用状况,都是通过一个统一的管理界面来监控的,因此维护变得简单且易操作。

3) 更多的访问和更好的竞争。公司员工不再需要通过本地网络来访问公司的信息,这就可以让公司员工甚至是合作商在任何地方访问他们所需要的数据。因为中小企业不再需要花费上千万美元来打造最新技术和最新应用以创造最好的系统,所以云存储为中小企业和大公司的竞争铺平了道路。事实上,云存储更有利于小企业,原因就是大企业已经花费重金打造了自己的数据存储中心。

云存储带来优势的同时,本身也潜藏着忧患与缺点:

1) 数据安全性。当所要存储的数据较为机密时,则会对将数据存放于云存储服务提供商的安全性产生疑虑。为了妥善地保护数据,来自于某一客户的数据必须与其他客户的数据适当地隔离开来。数据存储原来的地方,或是从一个地方移至其他的地方,都必须确保它们的安全。云服务提供商必须有相关的系统,以防止数据外泄或被第三方任意访问。适当的职责分权以确保审核与监控不会失效,即便是云服务提供商中有特权的用户也一样。

2) 数据访问性能。由于数据存储于云存储服务商的服务器上,数据访问要通过网络传输。访问性能可能比本地存储设备的性能低。数据的可靠性和可用性将取决于广域网,以及

服务商所提供的预防措施的好坏,一旦网络出现问题或服务器宕机,云端数据将无法访问。但好的云存储服务提供商是绝不允许这样的事情发生的,哪怕是0.001%的概率,也会给自身及客户带来巨大的损失。

根据用户的具体需求,云存储按类型可分为公共云存储、内部云存储、混合云存储,具体区别如下:

(1) 公共云存储

像亚马逊公司的 Simple Storage Service(S3) 和 Nutanix 公司提供的存储服务一样,它们可以低成本地提供大量的文件存储。供应商可以保证每个客户的存储、应用都是独立的、私有的。其中以 Dropbox 为代表的个人云存储服务是公共云存储发展较为突出的代表,国内比较突出的代表有搜狐企业网盘、百度云盘、乐视云盘、移动彩云、金山快盘、坚果云、酷盘、115 网盘、华为网盘、360 云盘、新浪微盘、腾讯微云、cStor 云存储等。

公共云存储可以划出一部分用作私有云存储。一个公司可以拥有或控制基础架构,以及应用的部署,私有云存储可以部署在企业数据中心或相同地点的设施上。私有云可以由公司自己的 IT 部门来管理,也可以由服务供应商管理。

(2) 内部云存储

这种云存储和私有云存储比较类似,唯一的不同点是它仍然位于企业防火墙内部。截至 2014 年可以提供私有云的平台有: Eucalyptus、3A Cloud、minicloud 安全办公私有云、联想网盘等。

(3) 混合云存储

这种云存储把公共云和私有云/内部云结合在一起。主要用于按客户的要求访问,特别是需要临时配置容量的时候。从公共云上划出一部分容量配置一种私有云或内部云使得公司能够面对迅速增长的负载波动或高峰。尽管如此,混合云存储提高了跨公共云和私有云分配应用的复杂性。

一般而言,云存储系统包括 4 个主要的部件,分别是:存储层、基础管理层、应用接口层和访问层。

(1) 存储层。存储层是云存储最基础的部分。存储设备可以是 FC 光纤通道存储设备,也可以是 NAS 和 iSCSI 等 IP 存储设备,还可以是 SCSI 或 SAS 等 DAS 方式的存储设备。云存储中的存储设备往往数量庞大且分布在不同的地域,彼此之间通过广域网、互联网或 FC 光纤通道网络连接在一起。存储设备之上是一个统一的存储设备管理系统,可以实现存储设备的逻辑虚拟化管理、多链路冗余管理,以及硬件设备的状态监控和故障维护。

(2) 基础管理层。基础管理层是云存储最核心的部分,也是云存储中最难以实现的部分。基础管理层通过集群、分布式文件系统和网格计算等技术,实现云存储中多个存储设备之间的协同工作,使多个存储设备可以对外提供同一种服务,并提供更高、更强、更好的数据访问性能。

(3) 应用接口层。应用接口层是云存储最灵活多变的部分。不同的云存储运营单位可以根据实际业务类型,开发不同的应用服务接口,提供不同的应用服务。比如视频监控应用平

台、IPTV 和视频点播应用平台、网络硬盘应用平台、远程数据备份应用平台等。

(4) 访问层。任何一个授权用户都可以通过标准的公共应用接口来登录云存储系统,享受云存储服务。云存储的运营单位不同,云存储提供的访问类型和访问手段也会有所不同。

6.5 本章小结

存储介质的不断发展直接推动了数据组织和管理的不不断发展,在纸带机时代,人们仅能手工地进行数据 I/O,并进行人工的数据组织与管理;在磁带年代,人们仅能顺序地访问数据;后来有了磁盘,人们可以顺序也可以随机地进行数据的访问;同这个发展轨迹并行的是人们的编程理念也开始发生变化——人们希望数据与程序能够分开;于是文件系统及文件管理系统走进了历史的舞台,不过文件管理系统自身有缺陷,如前文所说的一致性难以维护、冗余性难以回避等,为了有效地解决这些问题,计算机科学家发明了数据库技术并在后续至今的数十年里大行其道于几乎所有场景下的所有应用系统。数据库技术也从早期的层次型数据库发展到网络数据库,当然层次型数据库可以看作网络数据库的特例。再后来是关系型数据库,关系型数据库得以突飞猛进的发展除了其天生清晰的数据模型和良好的存取控制外,SQL 的发明绝对功不可没:SQL 把关系型数据库中晦涩的数学逻辑用类似自然语言的语法定义了数据操作及数据定义。虽然各关系型数据库厂商推出产品的时候会根据自己的产品特点增加或强化一些特征,但是在遵循 SQL 标准这方面都是一致的。SQL 太强大了,以至于它经常被混淆为关系型数据库的代名词,而事实上,SQL 只是一个访问关系型数据库、被认可了的标准化语言。

随着时代的进步和科技的发展,人类需要存储与管理的数据越来越大,特别是互联网时代的到来,互联网庞大的数据量和极高的峰值访问压力使得以增加内存、CPU 等节点性能的垂直伸缩方案 (Scale Up) 走入了死胡同,使用大量廉价的机器组建水平可扩展集群 (Scale Out) 成为绝大多数互联网公司的必然选择。这就意味着传统的集中数据管理思路必须因为时代的变迁演化为分布式文件管理,这是时代的必然,也是技术的无奈。用“无奈”这样的词来形容是因为分布式系统的 CAP 理论:在分布式数据组织与管理中,在数据的一致性、可用性和分区容忍性 3 个指标中,只能同时追求其中的两个指标,而必须舍弃三者中的一个。这样的无奈也再次表明了“没有免费午餐”的哲学精神。

随着数据管理与存储规模量级的不断扩大,传统的关系型数据库在应付大数据时代的挑战时逐渐显得捉襟见肘。也就是在这样的时机下,NoSQL 跃然出世。NoSQL 是一种非关系型数据库,NoSQL 以“键-值”对存储数据,天然地支持分布式,同时在顶层设计方面,NoSQL 非常关注数据高并发的读写和海量数据的存储等,NoSQL 在架构和数据模型方面做了“减法”的同时,在扩展和并发等方面做了“加法”,使得它天然就是为大数据而生的。因此在大数据大行天下的今天,NoSQL 的发展势头尤其猛烈,NoSQL 逐步蚕食传统数据库的市场份额已经成为不争的事实。当然 NoSQL 还存在着很多的不足,但是针对 NoSQL 的改进研究和

基于 NoSQL 为具体的场景设计并实现具体的解决方案几乎是一直在同步进行的。

或许是发展的必然,也或许是 NoSQL 发展的迅猛势头,由阿里巴巴于 2010 年提出并践行的“去 IOE”(其中最重要的是去“O”,此处的“O”是 Oracle,关系型数据库航母产品)倡议引发了各界的热议,棱镜门事件让一个纯粹的技术话题提高了国家对安全的重视(当然,这是有道理的),使得技术层面的“去 IOE”演变成为“去国外化”的倡议和行动。这对于国有厂商或许是一个非常好的机会,至少市场的天平在往国有化倾斜。不过更重要的是,国有厂商需要在这个难得的历史条件下在技术上有所突破。如果真的能够在技术上有革命性的突破,“去 IOE”就不再是争议了。

本章参考文献

- [1] Anderson J C, Lehnardt J, Slater N. CouchDB: the Definitive Guide [M]. Sebastopol: O' Reilly Media, Inc., 2010.
- [2] Chang F, Dean J, Ghemawat S, et al. Bigtable: A Distributed Storage System for Structured Data [J]. Proceedings of Usenix Symposium on Operating Systems Design & Implementation, 2006, 26(2): 205-218.
- [3] Chodorow K. MongoDB: The Definitive Guide [M]. Sebastopol: O' Reilly Media, Inc., 2013.
- [4] Han J, Haihong E, Le G, et al. Survey on NoSQL Database [C]. Pervasive Computing and Applications (ICPCA), 2011 6th international conference on. IEEE, 2011: 363-366.
- [5] Lakshman A, Malik P. Cassandra——A Decentralized Structured Storage System [J]. Operating Systems Review, 2010, 44(2): 35-40.
- [6] Sanfilippo S, Noordhuis P. Redis [EB/OL]. <http://redis.io>, 2010.
- [7] Sivasubramanian S. Amazon DynamoDB: A Seamlessly Scalable Non-relational Database Service [C]. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: 729-730.
- [8] Stonebraker M. SQL Databases v. NoSQL Databases [J]. Communications of the ACM, 2010, 53(4): 10-11.
- [9] Sumbaly R, Kreps J, Gao L, et al. Serving Large-scale Batch Computed Data with Project Voldemort [C]. Proceedings of the 10th USENIX Conference on File and Storage Technologies, 2012: 18-18.
- [10] Veltte T, Veltte A, Elsenpeter R. Cloud Computing, A Practical Approach [M]. New York: McGraw-Hill, Inc., 2009.
- [11] 任怡, 吴泉源, 贾焰, 等. 事务处理技术研究综述 [J]. 计算机研究与发展, 2005, 42(10): 1779-1784.
- [12] 石树刚. 关系数据库 [M]. 北京: 清华大学出版社, 1993.
- [13] 张俊华. 大数据日知录 [M]. 北京: 电子工业出版社, 2014.
- [14] 周江, 王伟平, 孟丹, 等. 面向大数据分析的分布式文件系统关键技术 [J]. 计算机研究与发展, 2014, 51(2): 382-394.

Chapter 7 第 7 章

数据表示与理解

本章的写作及润色,得到了南京大学计算机科学与技术系及智能信息处理研究组的张雷博士及彭岳、蒋澜、李红、王茜、陈厚兵、陆恒杨、蔡洋及王咏乾等几位同学的协助,在此表示深深的谢意。

7.1 引言

神说要有光,于是便有了光。

17 世纪初,笛卡儿提出光的微粒说和波动说两种假说,引发两种假说的世纪争论。

1655 年,格里马第首次提出“光的衍射”概念,其是光的波动学说的最早倡导者。

1663 年之后,波义耳及胡克陆续重复和提出类似格里马第的学说观点。

1666 年,惠更斯提出了光是一种机械波。

1672 年,牛顿利用三棱镜得到光谱,并开创了光的粒子学说。

1801 年,托马斯·杨设计了著名的杨氏双缝干涉,为光的波动性提供了有力的证据。

1818 年,菲涅耳在惠根斯学说的基础上提出波动的数学理论,并进行了实验。

1821 年,夫琅和费利用光栅研究了光的衍射现象。

1850 年,傅科用高速旋转镜法测出了光在真空中的传播速度。

1873 年,麦克斯韦全面建立起电磁场理论,并预言光是一种电磁波。

1887 年,赫兹发现了光电效应,光的粒子性再次得到证明。

1905 年,爱因斯坦提出了光的波粒二象性。

围绕光的本质问题,数百年来,不同学术派别的科学家们基于不同的科学范式(实验、理论、模拟)不断地展开尝试。正是他们的努力,才逐步揭开了掩盖在“光的本质”外面的那层扑朔迷离的面纱。从上面简述的历史中可以看出,每一次新的突破(包括或许是定论的波粒二象性)往往是基于已经发现的有别于前人所发现的反映光的本质的一些新属性,比如支持波动性观点的“干涉”“衍射”等特征属性,或者支持粒子性观点的“光电效应”“色谱”

等特征属性。

人们认识和理解目标主体行为、规律的关键在于对反映目标主体特征的认知和把握。在数据科学的大背景下,这意味着必须从原始数据中提取出对后续分析目标有贡献的特征来,才有可能发现隐藏在数据背后的知识和洞见。为了有效地达到这样的目标,我们势必要认真回答如下几个问题:

1) 数据表示是否支持和有助于后续的特征提取和选择。由于数据的来源和类型有很多,而不同数据源的数据表示和度量单位未必一致,这就意味着必须将不同量纲和度量意义的数据在一个统一的(评估)标准下进行度量,才能使得后续的数据处理,对于每一个数据源的数据都是公平的,这将涉及数据的规范表示问题。

“半斤八两”一般用来形容两样东西彼此一样,不相上下。其原因是:古时候中国的计量是十六进制的,因此半斤就是八两(宋·释惟白《建中靖国续灯录》中提及“踏着秤锤硬似铁,八两元来是半斤”)。显然,如果基于当今的度量进制“半斤=五两”来理解半斤八两就会产生误会。更进一步的探讨,当今的“1两”与古时的“1两”是一样的吗?答案自然是否定的,原因在于今时和古时的度量衡也是不一样的,文献表明:秦及西汉时期的“1两”折合为如今法定的计量是16.14克;而东汉及魏晋时期的“1两”折合为如今法定的计量为13.92克;北周时期的“1两”折合为如今法定的计量是15.66克……即便是离当今最接近的清代,彼时的“1两”折合为如今法定的计量是37.31克。上述这些数据表明,如果不在量纲上和进制上加以统一和规范,两个数字意义之间的比较是没有任何意义的,至少是大相径庭的。在数据处理中,也经常会遇到这样的场景。显然,不对数据进行规范化处理,其处理结果一定是有失偏颇的。

2) 数据表示是否支持,以及如何支持后续的特征提取和选择。由于数据的来源和类型有很多,因此,针对不同来源和类型的数据应该有不同的基本运算方法加以响应和应对,这将涉及数据度量方法的选型。

“君当作磐石,妾当作蒲苇。蒲苇韧如丝,磐石无转移”出自《孔雀东南飞》(汉乐府诗),描述的是刘兰芝和焦仲卿坚贞不二的爱情表白,此处不表。本文仅关注其中的比喻“蒲苇韧如丝”(其他雷同),蒲苇是禾本科蒲苇属下的一种植物,茎中空,末端有柔软、银色的毛;“韧”是一个形声字,本意是耐割、耐划的皮,转义为一种受外力作用时虽变形但不易折断的品质。

诗中将蒲苇的韧和丝(的韧)相比是建立在两者相似的基础上的(否则就谈不上比喻了),从数据分析的角度有几个问题需要回答:①蒲苇的韧和丝(的韧)相比有多大的相似程度。②蒲苇的哪些属性特征能够反映其在“韧”这个特点上的属性。第一个问题涉及两个对象(在“韧”这个方面)的相似度度量(显然如果两者相似性比较小,则意味着这种比喻是不恰当的);而第二个问题则需要在对蒲苇进行理性记录的时候要记录下能够反映分析(比较)目标的属性,而不是仅仅记录下其归属某科某属及其形状特征。这些特征对于评估其他

方面的指标或许有用,但对于当前场景的相似度比较目标而言还不够。

事实上,在任何数据分析的场景下,都会遇到类似的问题:如何记录和表示一个对象、如何从中提取对分析目标有贡献的特征、如何评估此特征与彼特征的相似度……所有这些问题都是数据分析的基础。

3) 原始数据中的哪些部分(成分)与后续的目标应用相关。这将涉及特征表示的话题,往往需要领域专家的经验、具体数据类型的常识认知作为支撑。如果在开展项目的过程中,没有任何的专家经验和常识认知作为支撑,那么原始数据就是特征。

图7-1描述的是一段人的脉象图示(中医专家就是通过将手指搭在人的寸关尺的位置触摸到类似的脉象),然后就可以从脉诊的角度进行脉象的识别(平、滑、涩……)或诊断出相应的证候(或病症)。

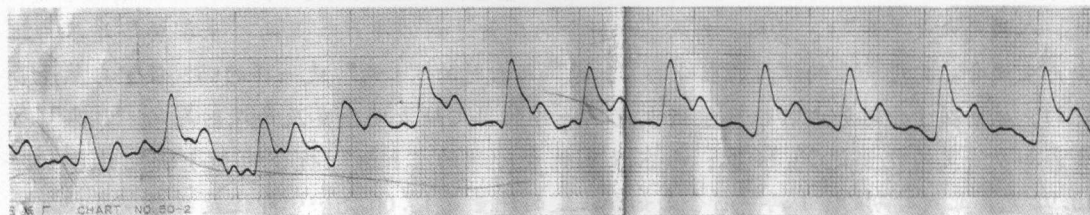


图7-1 脉象示意图

脉诊客观化研究的第一步是如何用计算机的手段进行脉象数据的采集,以及脉象的识别与诊断。前者的实现较为简单,麻烦点在于,从这样的周期图像中采集哪些特征会有助于后续的模式识别呢?中医专家从生理学的角度给出了一些有绝对借鉴价值的特征表示和提取建议,如图7-2所示。

如图7-2所示,中医专家认为, $(h_1, h_3, h_4, h_5, t_1, t_4, t_5, t, w)$ 是进行脉象研究的重要特征,因为这些特征都有其显然的生理学意义,比如:

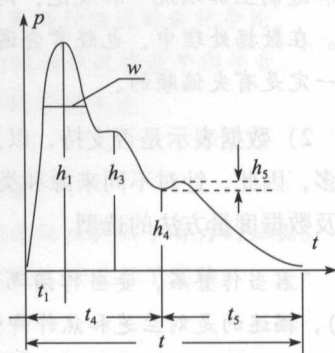


图7-2 脉象特征表示

1) h_1 的大小表示心脏射血的力度,这意味着 h_1 越大,射血力度越大。

2) h_3 的大小表示因为血管壁的弹性而进行的反弹,这意味着 h_3 越大,血管弹性越好。

基于专家的经验,从脉象采集序列中提取出一个完整的周期,然后提取出如此的特征,就可以进行脉诊的识别了(至于这些特征是否足够好或足够完备则是另外的话题,此处不表)。当然如果没有专家的经验支撑,整个脉象序列(或者一个脉象的一个周期)就是特征,或者在此特征基础上进行信号变换,提取出其对应的频率特征,这就耦合到后续的特征提取与特征选择了。

4) 从原始数据中提取出的特征是否是高质量的特征。这个问题指的是既有的特征表示对

后续的分析性能（精度、速度）的提升是否有贡献，这将涉及特征提取和特征选择两个问题，这两个问题的本质都是通过一系列的（非）线性映射或变换将原始特征映射到维度更低、质量更高的特征，或者从原始特征中按照某种评估指标挑选一些有助于后续分析的（低维）高质量特征。

身体不适，前往医院就诊，常用的检查指标包括体温、心率、血压等（中医使用“望闻问切”来获得中医理论支持的若干属性），事实上利用这些简单的指标就可以进行一些基本的判断。不过，如果觉得这些表象获取的数据还不够的话，还可以进行验血、验尿、X光、胸透、核磁等。事实上，每一种医疗检测手段都是从不同的角度来采集反映人不同生理指标的各项信息。所有的检测数据汇聚到一起就是反映人的整个生理体征的医疗数据（特征数据），医生需要根据这些特征数据对病人的疾病进行定性的诊断（识别过程）。显然，有经验的医生不需要对病人进行全方位的、几乎囊括医院所有检查能力的检查后才能给出“病毒性感冒”这样的诊断结论，而只会选择性地地进行某一个或有限几个（和可能的病症相关的）生理指标的检测，除了可以大范围降低病人的医疗成本之外，对于病症诊断而言，也是高效的。即便是将所有的检测数据汇聚到一起进行诊断，有经验的医生也会根据自身的医疗经验剔除所有检测数据之间的相关性，以及检测手段（仪器）可能存在的噪声并充分考虑各类检测数据与病情的相关性等，在一个凝练的特征属性中对病情进行诊断。

事实上，任何一个对象都是多元化的，而描述或记录一个对象的维度和方式也是多元化的。这就意味着：在进行数据分析时，应该提取（或选择）对后续分析目标有贡献的数据（或特征），这样在有效减少前期成本的同时也会对后期成本的降低大有裨益。

本章将顺序地介绍数据建模必须依赖的度量方法、数据规范和特征工程三大基础支撑，并对实际操作中的技术选型进行扼要的应用提示，本章后面的结构安排如下：7.2节介绍在数据分析场景下常用的度量方法，包括相似性度量和距离函数度量，前者尝试建立一个以两组数据为自变量的评估函数，函数的取值越大表示两者越相似；后者尝试构建一个以两组数据为自变量的评估函数，函数的取值越小表示两者越相似；7.3节介绍数据的规范化表示方法，借此保证所有的数据均都能公平地参与到或更加有利于后续的所有计算；7.4节从特征工程的角度介绍特征表示、特征提取和特征选择的基本概念，以及每一个概念下可能的经典技术选型；7.5节给出在实际应用中的若干建议；7.6节对本章进行小结。

7.2 度量方法

大千世界的万事万物（物理世界的模拟信号）能够被计算机记录、处理或分析的前提是万事万物能够被数字化、数据化，进而成为能够被存储的数据。数字化是指将模拟信号转化为数字信号的过程，包括采样和A/D转化两个过程，在计算机中用0和1表示，基本单元是位（bit）；数据化是将很多位加以结构化和颗粒化，形成标准化的、开放的、非线性的、通用

的数据,基本单元是字节(byte),其基本数据类型包括布尔型、整型、实数型、字符型、字符串型等。由这些基本的数据类型可以封装出更复杂的数据类型,比如结构体、类等。

普通非计算机工作者在计算机应用过程中熟悉的诸如文本、表格、网页、图像、视频等,在计算机工作者的视角里都是0和1的组合,只是用信息技术的手段把这些0和1以物理世界本来的形式表示了出来而已。

在计算机中,对上述所有数据的处理,本质上都是对0和1的加工和处理,是建立在数值计算基础上的处理,比如用1字节(8位)表示一个英文字符、2字节(16位)表示一个中文字符,而若干个英文字符或中文字符就组成了一段文本;用一个二维矩阵表示(每个元素是图像中对应位置中的颜色)一帧图像,而若干帧图像就组成了一段视频;音频可以用一段离散的数值来描述,其中每个数值表示对应时刻的声强。

由于文本、表格、网页、视频这样的数据具有数值意义以外的高级语义,因而针对这些数据的运算往往没有诸如针对数值计算那样“+”“-”“×”“÷”的简单算子。事实上,针对上述数据,大多数情况下,人们也不需要类似“+”“-”“×”“÷”的操作(有一种情况例外,两个图像进行“+”“-”操作是有物理语义的,比如一幅图相加意味着两个图像的叠加,这是智能视频处理中的一个基本操作,此处不议)。针对上述数据,人们更需要的操作是相似性度量,比如两幅图、两段音频或两段视频具有多大的相似性。如果上述问题得以回答,就可以进行更有意义的操作,比如:从很多图像中找出一组相似的图像(术语称为聚类,相似的图像属于一个聚集)。

相似性的度量方法有很多,有的适用于专门的领域,也有的适用于特定类型的数据,如何选择相似性的度量方法是一个相当复杂的问题。以聚类为例,刻画聚类数据之间的亲疏远近程度主要有相似系数函数和距离度量函数两类,以下将详细介绍。

7.2.1 相似系数函数

所谓相似系数函数,是指两个“数据”愈相似,则相似系数值愈大;两个“数据”愈不相似,则相似系数值愈小,往往取值范围设置在 $[0, 1]$ 或 $[-1, 1]$ 。这样就可以使用相似系数值来刻画数据的相似性,常用的相似系数函数有如下几种(包括但不限于):

(1) 夹角余弦

在几何学中,两个向量的夹角余弦(Cosine)可以表示两个向量的方向差异,如图7-3所示。向量 $A(x_1, y_1)$ 和 $B(x_2, y_2)$ 的夹角余弦是:

$$\cos\theta = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2}}$$

推广到 n 维空间上的两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 间的夹角余弦是:

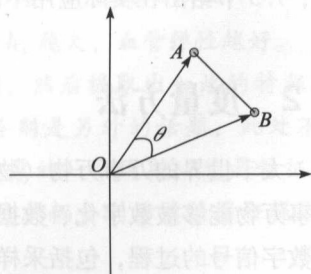


图7-3 夹角余弦示意

$$\cos\theta = \frac{\sum_{k=1}^n x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^n x_{1k}^2} \sqrt{\sum_{k=1}^n x_{2k}^2}}$$

由公式可知, 夹角余弦的取值范围为 $[-1, 1]$, 夹角余弦越大表示两个向量的夹角越小, 夹角余弦越小表示两向量的夹角越大, 当两个向量的方向完全重合时夹角余弦是 1, 当两个向量的方向完全相反时夹角余弦是 -1。

(2) 杰卡德相似系数

杰卡德相似系数 (Jaccard Similarity Coefficient) 是用于比较两个集合的相似度的指标, 所谓两个集合的杰卡德相似系数, 指的是两个集合 A 和 B 的交集元素在 A 和 B 的并集中所占的比例, 定义如下:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

在杰卡德相似系数的基础上衍生出一个杰卡德距离的概念, 定义如下:

$$J_*(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

杰卡德距离是用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度。

(3) 相关系数

相关系数 (Correlation Coefficient) 是衡量随机变量 X 与 Y 相关程度的一种方法, 相关系数的定义涉及了概率理论中数学期望的定义。对于离散型随机变量 X , 当 $X = X_i (i = 1, 2, \dots, n)$ 的概率为 $P = P_i (i = 1, 2, \dots, n)$, 则 X 的 (数学) 期望定义如下:

$$E[X] = \sum_{i=1}^n P_i X_i$$

X 的 (数学) 期望 $E[X]$ 是概率意义下的平均值。某工厂推销人员与工厂在货物物流方面的约定如下: ①将每箱货物按期无损地运到目的地可得到佣金 10 元。②如果不按期到目的地可得到佣金 8 元。③若货物有损则可得到佣金 5 元。④若既不按期又有损则得到佣金 -6 元 (相当于返款 6 元)。推销人员根据既有经验认为: ①货物按期无损地运到目的地的把握是 60%。②不按期到达目的地的可能是 20%。③货物有损的可能是 10%。④既不按期又有损的可能是 10%。上述的约定和推销人员的经验数据表示如表 7-1 所示。

表 7-1 数学期望计算示意

X	10	8	5	-6
P	0.6	0.2	0.1	0.1

表中, X (行) 表示约定的收益, P (行) 表示每一种收益的概率, 则推销人员在运送货物的时候, 每箱货物的期望获益是:

$$E[X] = \sum_{i=1}^n P_i X_i = 10 \times 0.6 + 8 \times 0.2 + 5 \times 0.1 - 6 \times 0.1 = 7.5$$

基于数学期望的定义, 随机变量 X 与 Y 的相关系数定义如下:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{D(X)} \sqrt{D(Y)}}$$

其中, 分母中的 $D(X)$ 、 $D(Y)$ 表示随机变量 X 与 Y 的方差 (方差的算术平方根称为标准差), 定义分别如下:

$$D(X) = \frac{\sum_{k=1}^n (X_k - \bar{X})^2}{n}$$

其中, \bar{X} 是随机变量 X 的平均值 ($D(Y)$ 的定义雷同于 $D(X)$, 在此不重复叙述)。

其中, 分子 $\text{Cov}(X, Y)$ 表示期望值分别为 $E[X]$ 和 $E[Y]$ 的两个随机变量 X 与 Y 的协方差 (期望是概率意义下的算术平均值, 见上文), 定义如下:

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

协方差表示的是两个变量总体误差的期望: 如果 X 与 Y 的变化趋势一致, 那么 $\text{Cov}(X, Y)$ 就是正值; 如果 X 与 Y 的变化趋势相反, 那么 $\text{Cov}(X, Y)$ 就是负值; 如果 X 与 Y 是统计独立的, 那么 $\text{Cov}(X, Y)$ 就是 0。

由相关系数的定义可知, 相关系数的取值范围是 $[-1, 1]$, 相关系数的绝对值越大, 则表明 X 与 Y 相关度越高, 当 X 与 Y 线性相关时, 相关系数取值为 1 (正线性相关) 或 -1 (负线性相关)。在相关系数的基础上, 相关距离 (Correlation Distance) 的定义如下:

$$d(X,Y) = 1 - \rho_{X,Y}$$

7.2.2 距离函数

距离函数是把每个数据都看作为高维空间中的一个点, 进而使用某种距离来表示数据之间的相似性。距离较近的样本点性质较相似, 距离较远的样本点则差异较大。一般用 $d(x, y)$ 来表示两个“数据” x 和 y 之间的距离, 显然 $d(x, y)$ 有很多种表达形式。

一般而言, 定义一个距离函数 $d(x, y)$ 需要满足以下几个准则:

$$1) d(x, x) = 0;$$

$$2) d(x, y) \geq 0;$$

$$3) d(x, y) = d(y, x);$$

$$4) d(x, k) + d(k, x) > d(x, y)。$$

上述准则中, 准则 1 约定了“自己到自己的距离为 0”; 准则 2 约定了“距离要非负”; 准则 3 约定了“距离度量是对称的”; 准则 4 被称为三角形法则, 即两边之和大于第三边。满足这 4 个条件的距离函数很多, 常用的距离函数有如下几种 (包括但不限于):

(1) 欧氏距离

欧氏距离 (Euclidean Distance) 是最易于理解的一种距离计算方法, 源自欧氏空间中两点间的距离公式。以两点为例, 二维平面上两点 $A(x_1, y_1)$ 和 $B(x_2, y_2)$ 间的欧氏距离是:

$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

推广到三维空间上两点 $A(x_1, y_1, z_1)$ 和 $B(x_2, y_2, z_2)$ 间的欧氏距离是:

$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

推广到 n 维空间上两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 间的欧氏距离是:

$$d(A, B) = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2}$$

(2) 曼哈顿距离

欧氏距离描述的是两点之间的距离。但在很多情况下,比如在一个城市的街区,从一个十字路口 $A(x_1, y_1)$ 开车到另外一个十字路口 $B(x_2, y_2)$, 如图 7-4 所示, 假设无法穿越直线距离中的障碍物, 那么两点之间的距离就无法用欧氏距离来计算了。针对这种场景, 专门引入了另外一种距离度量, 即曼哈顿距离 (这个问题的原型最早来源于曼哈顿岛的街区距离度量, 故以“曼哈顿”称之)。

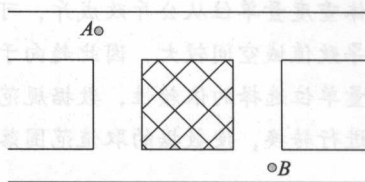


图 7-4 曼哈顿距离示意

两个点 $A(x_1, y_1)$ 和 $B(x_2, y_2)$ 之间的曼哈顿距离 (Manhattan Distance) 定义如下:

$$d(A, B) = |x_1 - x_2| + |y_1 - y_2|$$

推广到 n 维空间上两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 间的曼哈顿距离是:

$$d(A, B) = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

(3) 切比雪夫距离

上述的曼哈顿距离基于从 A 点到 B 点的转移过程中 (以二维平面为例), 每个步骤或者沿 X 方向, 或者沿 Y 方向。但是在有些场合, 比如国际象棋中, 国王走一步能够移动到相邻的 8 个方格中的任意一个, 那么国王从格子 $A(x_1, y_1)$ 走到格子 $B(x_2, y_2)$ 最少需要 $\max(|x_2 - x_1|, |y_2 - y_1|)$ 步, 一种类似的距离度量方法称为切比雪夫距离 (Chebyshev Distance), 二维平面中任意一个格子 $A(x_1, y_1)$ 走到格子 $B(x_2, y_2)$ 的切比雪夫距离是:

$$d(A, B) = \max(|x_2 - x_1|, |y_2 - y_1|)$$

推广到 n 维空间上两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 间的切比雪夫距离是:

$$d(A, B) = \max_i (|x_{2i} - x_{1i}|)$$

(4) 汉明距离

汉明距离 (Hamming Distance) 一般用于比较两个等长字符串 s_1 与 s_2 之间的编辑距离, 即将其中一个变为另外一个所需要进行的最小替换次数。例如字符串 “1111” 与 “1001” 之间的汉明距离为 2, 而字符串 “1011” 与 “1001” 之间的汉明距离为 1。

7.3 数据规范

7.2 节简单介绍了若干种常用的数据度量方法。在具体的应用场景下往往需要进行若干的改进（更多情况下，可以直接使用），或者对度量方法进行改进，或者对数据进行加工，数据规范是其中必不可少的一个环节。

数据的度量单位往往会影响到数据分析，比如把人的身高度量单位从米变成厘米、把人的体重度量单位从公斤改成斤，可能会导致完全不同的结果。一般而言，用较小的单位表示将导致值域空间较大，因此趋向于使这样的属性具有较大的影响或较高的权重。为了避免对度量单位选择的依赖性，数据规范化（或标准化）是一个解决方案，其基本的思路是：将数据进行转换，使数据的取值范围落入较小的一个共同区间，比如 $[-1, 1]$ 或 $[0, 1]$ 。

所谓数据规范化（Normalization）就是将数据按比例缩放，使之落入一个较小的特定区间。在某些比较和评价的指标处理中经常会用到，去除数据的度量单位限制，将其转化为无量纲的纯数值，以便于不同单位或量级的指标进行比较和加权。

数据规范化试图赋予数据不同维度的属性以相等的权重，借此让上述的度量方法更加合理和有效。数据规范化的主要作用（和优势）在于既可以保持数据的完整性，又可以最小化数据的冗余。数据规范化处理主要包括数据同趋化处理和无量纲化处理两个方面。前者主要解决不同性质数据的问题，这是因为对不同性质的指标直接（度量）分析不能正确地反映不同维度的作用力的综合结果，须先考虑改变不同维度的数据性质，使所有维度对度量结果的作用力同趋化；后者主要解决数据的可比性。数据标准化的方法有很多种（包括但不限于）：

（1）最小-最大规范化

所谓最小-最大规范化（Min-Max Normalization），也称为离差标准化，是对原始数据的线性变换，使结果值映射到 $[0, -1]$ 之间，对于给定的样本 $X = \{x_1, x_2, \dots, x_n\}$ ，经过最小-最大规范化以后得到 $X^* = \{x_1^*, x_2^*, \dots, x_n^*\}$ ，变换规则如下：

$$x_i^* = \frac{x_i - \text{Min}(X)}{\text{Max}(X) - \text{Min}(X)}$$

其中， $\text{Max}(X)$ 和 $\text{Min}(X)$ 分别代表样本 $X = \{x_1, x_2, \dots, x_n\}$ 中的最大值和最小值，该方法的缺陷在于：当有新数据加入时，可能会导致 $\text{Max}(X)$ 和 $\text{Min}(X)$ 发生变化，需要重新计算。

（2）Z 分数规范化

所谓 Z 分数规范化（Z-score Normalization）指的是利用均值和标准差对样本 $X = \{x_1, x_2, \dots, x_n\}$ 中的每个元素进行规范化，使经过处理的数据符合标准正态分布。对于给定的样本 $X = \{x_1, x_2, \dots, x_n\}$ ，经过 Z 分数规范化以后得到 $X^* = \{x_1^*, x_2^*, \dots, x_n^*\}$ ，变换规则如下：

$$x_i^* = \frac{x_i - \mu}{\sigma}$$

其中 μ 和 σ 分别表示样本 $X = \{x_1, x_2, \dots, x_n\}$ 的均值和标准差,该方法的缺陷在于:当有新数据加入时,如果新数据的取值落在最小最大范围之外,则该方法将面临“越界”错误。

标准欧氏距离 (Standardized Euclidean Distance) 就是基于Z分数规范化的思想对前文提及的欧氏距离的改良方法,基本改良思路是:假定数据各维分量的分布不一样,将各个分量都“标准化”到均值、方差相等的状态。

按照Z分数规范化的操作,将两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 分别转换为 $A^*(x_{11}^*, x_{12}^*, \dots, x_{1n}^*)$ 和 $B^*(x_{21}^*, x_{22}^*, \dots, x_{2n}^*)$,其中:

$$x_{1k}^* = \frac{x_{1k} - \mu_k}{\sigma_k}$$

$$x_{2k}^* = \frac{x_{2k} - \mu_k}{\sigma_k}$$

μ_k 和 σ_k 分别表示两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 各个维度分量的均值和标准差。

两点 $A^*(x_{11}^*, x_{12}^*, \dots, x_{1n}^*)$ 和 $B^*(x_{21}^*, x_{22}^*, \dots, x_{2n}^*)$ 间的欧氏距离是:

$$d(A^*, B^*) = \sqrt{\sum_{k=1}^n (x_{1k}^* - x_{2k}^*)^2} = \sqrt{\sum_{k=1}^n \left(\frac{x_{1k} - x_{2k}}{\sigma_k} \right)^2}$$

即两点 $A(x_{11}, x_{12}, \dots, x_{1n})$ 和 $B(x_{21}, x_{22}, \dots, x_{2n})$ 之间的标准化欧式距离是:

$$d(A, B) = \sqrt{\sum_{k=1}^n \left(\frac{x_{1k} - x_{2k}}{\sigma_k} \right)^2}$$

(3) 按小数定标规范化

所谓按小数定标规范化 (Decimal Scaling Normalization), 指的是通过移动样本 $X = \{x_1, x_2, \dots, x_n\}$ 中的每个元素的小数点位置进行规范化, 小数点的移动位数依赖于样本 $X = \{x_1, x_2, \dots, x_n\}$ 中的最大绝对值 $\text{Max}(|X|)$ 。对于给定的样本 $X = \{x_1, x_2, \dots, x_n\}$, 经过按小数定标规范化以后得到 $X^* = \{x_1^*, x_2^*, \dots, x_n^*\}$, 变换规则如下:

$$x_i^* = \frac{x_i}{10^j}$$

其中, j 是使得 $\text{Max}(x_i^*) \leq 1$ 的最小整数。

7.4 特征工程

特征工程是机器学习应用的基础,指的是利用领域知识从原始数据中提取特征(向量)的过程,以用于后续机器学习及数据挖掘。整个过程涉及诸如特征表示、特征提取、特征选择等几个基本内容(概念)。

7.4.1 特征表示

所谓特征表示,就是将数据转化为有利于后续分析和处理的形式而进行的一种形式化表示和描述,隐含着如下几个要素:

- 1) 特征表示的研究对象是原始数据,而原始数据具有不同的类型,如文本、音频、图像、视频等,这就意味着针对不同的数据类型应该使用不同的特征表示方法。
- 2) 特征表示的研究目标是有利于后续分析和处理,而后续的分析 and 处理都是有应用目标导向的,这就意味着在进行特征表示的时候必须要考虑后续应用的目标。
- 3) 特征表示的最终输出是可计算的特征向量,而此特征向量应该能如实、无歧义地表征原始数据在应用目标指向上的属性特征。
- 4) 对于给定的原始数据,在进行特征表示的相关研究和应用实践时,往往需要领域专家的知识 and 经验,基于专家经验提取的特征往往具有一定的物理意义。
- 5) 鉴于存放在计算机中的原始数据本身都已数字化,这就意味着,原始数据本身就是一种表示描述对象的特征向量。

在实际应用中,特征表示往往与下文依次介绍的特征提取、特征选择联合使用。

7.4.2 特征提取

特征提取,也称为特征抽取 (Feature Extraction),指的是从原始特征 $X = (x_1, x_2, \dots, x_N)$ 中重构出一组新特征 $Y = (y_1, y_2, \dots, y_M)$ 的过程,其数学描述为 $Y = f(X)$, 其中 $f(\cdot)$ 为重构函数。在实际应用中,特征提取往往与 7.4.3 节介绍的特征选择联合使用。值得一提的是:特征提取的过程是“ $X \rightarrow Y$ ”的降维转换(映射)过程,这个过程未必是可逆的,这就意味着从 Y 未必能够不失真地恢复回 X 。在机器学习应用中,之所以使用特征提取后的降维特征数据而不是利用原始的特征数据,或许有如下几个原因(包括但不限于):

- 1) 在原始的高维特征向量空间中,往往包含有冗余信息及噪声信息,这对于后续分析的准确率不利,而通过降维,有望减少冗余信息所造成的误差,从而提高后续分析的精度。
- 2) 仅在变量层面上分析可能会忽略变量之间的潜在联系,而通过降维,加速后续计算速度的同时,往往能够得到数据内部的结构特征。
- 3) 高维空间本身具有稀疏性(一维正态分布有 68% 的值落于正负标准差之间,而在十维空间上则只有 0.02%),通过降维,也能有效地解决原始数据中的稀疏性(sparse)问题。

总之,降维的目的在于减少预测变量个数的同时,确保这些变量是相互独立的,并能够提供一框架来解释结果。常见的特征提取方法有:主成分分析(PCA)、线性判别分析(LDA)、独立分量分析(ICA)、粗糙集属性约简等,以下简单介绍。

1. PCA

PCA(Principal Component Analysis, 主成分分析)是最常用的线性降维方法(降维的同时

也能够去除噪声),其目标是把原先的 N 个特征用数目更少的 M 个特征来取代,新特征是旧特征的线性组合,这些线性组合将最大化样本方差,尽量使新的 M 个特征互不相关。

PCA 是通过某种线性投影,将高维的数据映射到低维的空间中来表示,并期望在所投影的维度上数据的方差最大,以此既可使用较少的数据维度,同时又能保留住较多的原数据点的特性。通俗的理解,如果把所有的点都映射到一起,那么几乎所有的信息(如点和点之间的距离关系)都丢失了,而如果映射后方差尽可能的大,那么数据点就会分散开来,以此来保留更多的信息。可以证明:PCA 是丢失原始数据信息最少的一种线性降维方式,但是 PCA 并不试图去探索数据内在的结构。

样本 X 和样本 Y 的协方差 (Covariance) 定义如下:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

其中, \bar{X} 和 \bar{Y} 表示样本 X 和样本 Y 的平均值, N 表示样本 X 和样本 Y 的元素个数。如果两个样本的协方差为正时,则说明样本 X 和样本 Y 是正相关关系;如果两个样本的协方差为负时,则说明样本 X 和样本 Y 是负相关关系;如果两个样本的协方差为 0 时,则说明样本 X 和样本 Y 相互独立。显然, $\text{Cov}(X, Y) = \text{Cov}(Y, X)$, 同时,样本 X 和样本 X 的协方差即为样本 X 的方差,即

$$\text{Cov}(X, X) = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1} = \text{var}(X)$$

其中 $\text{var}(X)$ 表示样本 X 的方差。

当样本是 n 维数据时,它们的协方差就是协方差矩阵(对称方阵),方阵的边长是 C_n^2 ,比如对于三维数据 (x, y, z) ,计算它的协方差就是:

$$C = \begin{pmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) & \text{Cov}(x, z) \\ \text{Cov}(y, x) & \text{Cov}(y, y) & \text{Cov}(y, z) \\ \text{Cov}(z, x) & \text{Cov}(z, y) & \text{Cov}(z, z) \end{pmatrix}$$

若 $AX = \lambda X$, 则称 λ 是 A 的特征值, X 是对应的特征向量。基于这样的定义,可以理解为:矩阵 A 作用在它的特征向量上,仅仅使得 X 的长度发生了变化,缩放比例就是相应的特征值 λ 。

当 A 是 n 阶可逆矩阵时, A 与 $P^{-1}AP$ 相似,相似矩阵具有相同的特征值。特别地,当 A 是对称矩阵时, A 的奇异值等于 A 的特征值,存在正交矩阵 $Q(Q^{-1} = Q^T)$,使得:

$$Q^T A Q = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

对 A 进行奇异值分解就能求出所有特征值和 Q 矩阵。

$$A * Q = Q * D$$

其中, D 是由特征值组成的对角矩阵, 由特征值和特征向量的定义可知, Q 的列向量就是 A 的特征向量。

基于上述定义和相关介绍, PCA 的算法流程描述如下:

输入: 原始矩阵 $A_{N \times M}$

输出: 降维矩阵 $A_{N \times M'}$

Step1: 特征中心化: 将矩阵 A 中每一维的数据都减去该维的均值, 得到矩阵 B 。这里的“维”指的是一个特征 (或属性), 变换之后每一维的均值都变成了 0。

Step2: 计算矩阵 B 的协方差矩阵 C 。

Step3: 计算协方差矩阵 C 的特征值 S 和特征向量 V , $C = V * S * V^{-1}$ 。

Step4: 将特征值由大到小排列, 选择前 M' 个特征值对应的特征向量 (通过样本点中心化矩阵的相邻奇异值之间的比值大小, 或者采用特征值所占百分比, 例如大于 85%, 的方法来确定 M' 的大小), 得到一个 $M \times M'$ 的矩阵 X 。令 $A'_{N \times M'} = A_{N \times M} X_{M \times M'}$, 这样我们就把 $N \times M$ 的原始数据集 A 映射成了 $N \times M'$ 的数据集 A' , 特征由 M 个减到了 M' 个。

PCA 的优点是简单且无参数限制, 具有普适性, 因此应用极其广泛, 最重要的应用是对原有数据进行简化, 能有效地找出数据中最“主要”的元素和结构, 去除噪声和冗余, 将原有的复杂数据降维, 揭示隐藏在复杂数据背后的简单结构。但 PCA 模型本身也存在诸多的假设条件, 从而决定它存在一定的限制, 在有些场合可能会造成效果不好甚至失效。

PCA 的几个主要假设:

1) 线性: PCA 的内部模型是线性的, 这也就决定了它能进行的主元分析之间的关系也是线性的, 而在实际情况下未必如此, 现在比较流行的 kernel-PCA 的一类方法就是使用非线性权的值对原有 PCA 技术的拓展。

2) 大方差对应重要数据结构: PCA 隐含的一个强假设是原数据具有很高的信噪比, 因此大方差对应重要的数据结构, 而方差小则对应噪声, 但实际情况下可能未必如此。

3) 主成分之间正交: 这个假设使得 PCA 的求解可以采用线性代数分解的技术来实现, 如特征值分解和 SVD。

PCA 基于上述假设及 PCA 算法本身的特点导致其具有一些固有的缺点 (包括但不限于):

1) 当样本点具有一些非线性性质时, 采用 PCA 得到的降维结果无法反映出样本点之间所隐藏的非线性性质。

2) PCA 能找到很好地代表所有样本点的方向, 但这个方向对于后续的分析 (比如分类) 未必是最有利的。

3) 对 PCA 所要保持的主分量的个数的估计比较困难, 虽然有策略和方法来确定主分量的个数, 但是当奇异值的大小变化比较平缓时, 难以估计应该舍弃哪些分量。

4) 有些情况下, 难以对 PCA 所保持的主分量的意义进行解释。

2. LDA

LDA (Linear Discriminant Analysis, 线性判别分析), 也称为 Fisher 线性判别 (Fisher Linear Discriminant, FLD), 其基本思想是: 将高维的模式样本投影到最佳鉴别向量空间, 以达到抽取分类信息和压缩特征空间维数的效果。投影后保证模式样本在新的子空间中有最小的类内距离和最大的类间距离, 即模式在该空间中有最佳的可分离性。因此, 它是一种有效的特征提取方法。

LDA 与 PCA 都是常用的降维方法, PCA 聚焦于特征的协方差, 尝试找到比较好的投影方式; 而 LDA 更多的是考虑标注, 即希望投影后不同类别之间数据点的距离更大, 同一类别的数据点更紧凑。

以一个简单的实例来解释 LDA 的基本思想, 如图 7-5 所示。图 7-5 中的两幅图分别表示了对两个类 (C_1 、 C_2) 的不同投影方案所得到的不同结果。从图 7-5 的左图可以发现, 无论是在 x 轴还是 y 轴投影, 两个不同的类在投影轴上均有重叠的部分 (导致分类效果变差); 而从图 7-5 的右图可以发现, 如果能够找到 (新生成) 一个坐标轴 x' (投影函数), 将 C_1 和 C_2 投影到坐标轴 x' 上, 在此坐标轴上就可以把两个类很清晰地分开 (类间距离大、类内距离小, 有助于分类)。LDA 算法就是寻找这个 x' 轴 (投影函数) 的方法, 这个 x' 轴 (投影函数) 也称为 Fisher 线性判别。

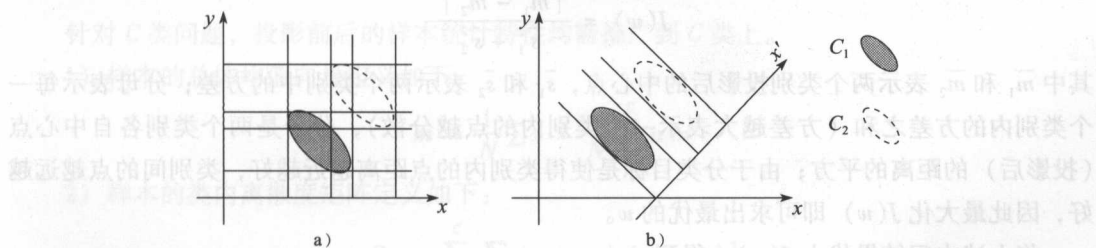


图 7-5 LDA 基本思想

LDA 本质上是一个线性分类器, 对于一个 k -分类问题, 需要有 k 个线性函数 (实际操作中, 只需要 $k-1$ 个线性函数即可):

$$y_k(x) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

其中, \mathbf{w}_k^T 称为权向量 (weight vector) 或法向量 (normal vector), w_{k0} 称为阈值 (threshold) 或偏置 (bias)。对于每一个分类, 都有一个公式计算其分值, 在所有的分类中, 分值最大的那个就是所属的分类。

对于一个二分类问题, 区别两类的投影函数 (直线) 为:

$$y(x) = \mathbf{w}_1^T \mathbf{x}$$

LDA 分类的一个目标是使得不同类别之间的距离越远越好, 同一类别之中的距离越近越好, 因此, 需要定义如下几个关键的度量。

1) 类别 i 的原始中心点 (均值) 为:

$$m_i = \frac{1}{N_i} \sum_{x \in D_i} x$$

其中, i 表示类别, N_i 是类别 i 中的元素个数, D_i 表示类别 i 中的元素 (点)。

2) 类别 i 投影后的中心点为: $\bar{m}_i = w^T m_i$

3) 类别 i 中投影后各个点之间的方差为:

$$\bar{s}_i = \sum_{y \in Y_i} (y - \bar{m}_i)^2$$

其中, Y_i 表示类别 i 中各个点投影后各自对应的目标值, \bar{s}_i 用于评估类别 i 投影后, 类别点之间的分散程度。将 \bar{m}_i 代入 \bar{s}_i 可得到:

$$\bar{s}_i = \sum_{x \in D_i} (w^T x - w^T m_i)^2 = \sum_{x \in D_i} w^T (x - m_i) (x - m_i)^T w$$

4) 令

$$S_i = \sum_{x \in D_i} (x - m_i) (x - m_i)^T$$

其中, D_i 表示类别 i 中的元素 (点), 则上式可以化简为:

$$\bar{s}_i = w^T S_i w$$

5) LDA 投影后的目标优化函数是:

$$J(w) = \frac{|\bar{m}_1 - \bar{m}_2|^2}{\bar{s}_1^2 + \bar{s}_2^2}$$

其中 \bar{m}_1 和 \bar{m}_2 表示两个类别投影后的中心点, \bar{s}_1 和 \bar{s}_2 表示两个类别中的方差; 分母表示每一个类别内的方差之和 (方差越大表示一个类别内的点越分散), 分子是两个类别各自中心点 (投影后) 的距离的平方; 由于分类目标是使得类别内的点距离越近越好, 类别间的点越远越好, 因此最大化 $J(w)$ 即可求出最优的 w 。

将上述中间结果代入 $J(w)$, 得到:

$$J(w) = \frac{w^T (m_1 - m_2) (m_1 - m_2)^T w}{w^T S_1 w + w^T S_2 w} = \frac{w^T S_B w}{w^T S_w w}$$

其中 $S_B = (m_1 - m_2) (m_1 - m_2)^T$, $S_w = S_1 + S_2$ 。

最大化 $J(w)$ 需对 $J(w)$ 按变量 w 求导并使 $\frac{\partial J(w)}{\partial w} = 0$, 即

$$\frac{\partial J(w)}{\partial w} = \frac{\partial \left(\frac{w^T S_B w}{w^T S_w w} \right)}{\partial w} = \frac{S_B w (w^T S_w w) - S_w w (w^T S_B w)}{(w^T S_w w)^2} = 0$$

则需 $S_B w (w^T S_w w) - S_w w (w^T S_B w) = 0$, 将 $J(w) = \frac{w^T S_B w}{w^T S_w w}$ 代入, 得

$$S_B w = J(w) S_w w$$

令 $\lambda = J(w)$, 则

$$S_B w = \lambda S_w w$$

这是一个广义特征值问题, 若 S_w 非奇异, 则

$$S_w^{-1} S_B w = \lambda w$$

因此可通过对 $S_w^{-1} S_B$ 特征值分解, 将最大特征值对应的特征向量作为最佳投影方向 w , 以下给出 w 的推导思路。

因为 $S_B = (m_1 - m_2)(m_1 - m_2)^T$, 所以

$$S_B w = (m_1 - m_2)(m_1 - m_2)^T w = (m_1 - m_2) \cdot \lambda_w$$

其中, $\lambda_w = (m_1 - m_2)^T w$, 是一标量, 将上式左乘 S_w^{-1} , 得到:

$$S_w^{-1} S_B w = S_w^{-1} (m_1 - m_2) \cdot \lambda_w$$

将 $S_w^{-1} S_B w = \lambda w$ 代入上式, 得到:

$$S_w^{-1} (m_1 - m_2) \cdot \lambda_w = \lambda \cdot w$$

因此, $w = S_w^{-1} (m_1 - m_2) \lambda_w / \lambda$, 由于特征向量仅与向量方向相关, 故

$$w = S_w^{-1} (m_1 - m_2)$$

以上 Fisher 准则只能用于解决二分类问题, 对于 C 类问题, 则需要用 $C-1$ 个上述的二分类的 Fisher 线性判别函数, 即需要由 $C-1$ 个投影向量 w 组成的投影矩阵 $W \in R^{d \times (C-1)}$, 将样本投影到此投影矩阵上, 从而可以提取 $C-1$ 维的特征向量 $y \in R^{C-1}$ 。

$$y = W^T x$$

针对 C 类问题, 投影前后的样本统计特性均需推广到 C 类上。

1) 样本的总体均值向量定义如下:

$$m = \frac{1}{N} \sum x = \frac{1}{N} \sum_{i=1}^C n_i m_i$$

2) 样本的类内离散度矩阵定义如下:

$$S_w = \sum_{i=1}^C \sum_{x \in X_i} (x - m_i)(x - m_i)^T$$

3) 样本的类间离散度矩阵定义为:

$$S_B = \sum_{i=1}^C n_i (m_i - m)(m_i - m)^T$$

4) 投影后样本均值向量定义如下:

$$\bar{m} = \frac{1}{N} \sum y = \frac{1}{N} \sum_{i=1}^C n_i \bar{m}_i$$

5) 投影后样本的类内离散度矩阵:

$$\bar{S}_w = \sum_{i=1}^C \sum_{y \in Y_i} (y - \bar{m}_i)(y - \bar{m}_i)^T$$

6) 投影后样本的类间离散度矩阵:

$$\bar{S}_B = \sum_{i=1}^C n_i (\bar{m}_i - \bar{m})(\bar{m}_i - \bar{m})^T$$

基于上述定义, Fisher 准则也可推广到 C 类问题:

$$J(w) = \frac{\bar{S}_B}{S_w} = \frac{W^T S_B W}{W^T S_w W}$$

为了使 Fisher 准则取得最大值, 类似二分类问题, W 需要满足:

$$S_B W = \lambda S_w W$$

如果 S_w 非奇异, 则 W 的每一列都是 $S_w^{-1} S_B$ 的前 $C-1$ 个较大特征值对应的特征向量。

以上简单地介绍了 LDA 的基本思想和推导过程, 应当注意到:

1) LDA 降维是直接和类别的个数相关的, 与数据本身的维度没关系, 比如原始数据是 N 维的, 共有 C 个类别, 那么 LDA 降维之后, 其维度在 $(1, C-1)$ 之间; 这一点与 PCA 截然不同, PCA 降维是直接和数据维度相关的, 比如原始数据是 N 维的, 那么 PCA 后, 其维度在 $(1, N-1)$ 之间。

2) LDA 根据类别的标注进行降维, 关注分类能力, 因此不保证投影到的坐标系是正交的 (一般都不正交); 与之截然不同是 PCA, PCA 投影的坐标系都是正交的。

3) LDA 投影后最多只能保留 $C-1$ 维, 可能对于一些问题来说, 特征数目太少。

4) LDA 的本质是参数估计方法, 假设分布是符合单峰的高斯分布, 对于不符合此假设条件的数据集, 则没法保留标签信息。

5) 对于那些由方差而不是均值来区分的数据来说, LDA 同样无法处理。

3. ICA

ICA (Independent Component Analysis) 是 20 世纪 90 年代朱顿 (Jutten) 和埃罗 (Herault) 提出的一种将观察到的数据进行某种线性分解, 使其分解成统计独立的成分的技术。从统计的角度来看, ICA 和 PCA 同属于多变量数据分析方法, 但 ICA 处理得到的各个分量不仅去除了相关性, 还是相互统计独立且非高斯分布的。因此, ICA 能够更加全面地揭示数据间的本质结构。

“鸡尾酒会问题” (cocktail party problem) 是计算机语音识别领域的一个问题, 当前的语音识别技术已经可以以较高的精度识别一个人所讲的话, 但是当说话的人数为两人或多人时, 语音识别率就会极大地降低, 这一难题被称为鸡尾酒会问题。问题原型描述如图 7-6 所示, 其中 $s_i(t)$ ($i=1, 2, 3$) 是声音源, $x_i(t)$ ($i=1, 2, 3$) 是放置在三个不同地方的麦克风录制的声音, 显然, 每一个麦克风录制的声音都是三个声音源的叠加 (声音是独立的且忽略声音的延迟), 距离的远近对叠加的影响是不一样的, 权重用 a_{ij} ($i=1, 2, 3; j=1, 2, 3$) 表示, 则

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{21}s_2(t) + a_{31}s_3(t) \\ x_2(t) = a_{12}s_1(t) + a_{22}s_2(t) + a_{32}s_3(t) \\ x_3(t) = a_{13}s_1(t) + a_{23}s_2(t) + a_{33}s_3(t) \end{cases}$$

其中权重矩阵可以表示为:

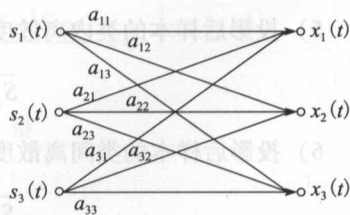


图 7-6 鸡尾酒会问题原型

$$A = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

对于“鸡尾酒会问题”，如果知道了权重矩阵 A ，通过观测数据 $x_i(t)$ ($i=1, 2, 3$) 是能够反推出 $s_i(t)$ ($i=1, 2, 3$) 的。但实际的问题在于：权重矩阵 A 往往

不可知，ICA 就是尝试在权重矩阵 A 不可知的情况下进行的一种估计方法。ICA 模型的基本示意图如图 7-7 所示。

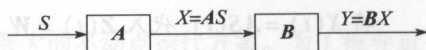


图 7-7 ICA 模型

如图 7-7 所示，矩阵 A 称为混合矩阵，输入 S ，输出 X ，ICA 模型的目标是找出解混矩阵 B ，然后通过此解混矩阵计算出 Y （原始信号的估计）。ICA 的基本假设和约定有：

1) 独立成分被假定是统计独立的，该假设是 ICA 能够成立的前提。所谓随机变量 y_i ($i=1, 2, \dots, N$) 统计独立，是指在 $i \neq j$ 时，有关 y_i 的取值情况对于 y_j 如何取值没有提供任何信息。

2) 独立成分具有非高斯的分布，这是一个强假设，因为如果观测到的变量具有高斯分布，那么 ICA 在本质上就不可能实现。这是因为：如果 S 经过混合矩阵 A 后，独立成分联合概率不发生变化的话，就无法从混合中得到混合矩阵的信息。

3) 假定混合矩阵是方阵，即独立成分的个数和观测到的混合量个数相同。

ICA 算法的基本步骤为：

1) 标准化：数据标准化的主要目的是从观测数据中除去其均值；具体操作雷同 PCA 中的操作，即将矩阵 X 中的每一维数据都减去该维的均值，得到矩阵 X' 。这里的“维”指的是一个特征（或属性），变换之后每一维的均值都变成了 0。

2) 白化：也称为球化，其主要目的是去除数据的相关性。数据的白化处理可以使随后的计算大为简化，并且还可以压缩数据。白化过程一般包括两个基本步骤，分别是：先对 X' 进行主成分分析，然后再对 X' 进行白化。

① 计算矩阵 X' 的协方差矩阵 C ，然后计算 C 的特征值 Λ 和特征向量 U ，即

$$C = UX'U^{-1}$$

② 对 X 进行白化。

对观测信号 $X(t)$ ，通过线性变换将 $X(t)$ 投影到新的子空间后变成白化向量，即

$$Z(t) = W_0 X(t)$$

其中， W_0 为白化矩阵， $Z(t)$ 是白化向量， W_0 通过 X' 的协方差矩阵的特征值 S 和特征向量 V 进行一个变换，变换的方法是：

$$W_0 = \Lambda^{-\frac{1}{2}} U^T$$

白化的目的是去除数据的相关性，所谓白化向量指的是：对于一个零均值的向量 Z ，如果 $E\{ZZ^T\} = I$ ，其中 I 是单位向量，则称此向量 Z 为白化向量。 W_0 满足白化变换的需求，以下

将简单证明。对于向量 $Z(t) = \Lambda^{-\frac{1}{2}} U^T X(t)$, 其协方差矩阵:

$$E\{ZZ^T\} = E\{\Lambda U^T X X^T U \Lambda^{-\frac{1}{2}}\} = \Lambda^{-\frac{1}{2}} U^T E\{X X^T\} U \Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}} = I$$

故 W_0 满足白化变化的需求。

将 $X(t) = AS(t)$ 代入 $Z(t) = W_0 X(t)$, 并令 $\bar{A} = W_0 A$, 则

$$Z(t) = W_0 AS(t) = \bar{A}S(t)$$

在 ICA 中, 对于零均值的独立源信号 $S(t)$ 而言, 有

$$E\{S_i S_j\} = E\{S_i\} E\{S_j\} = 0 \quad (i \neq j)$$

其中, 协方差矩阵是单位阵, 即 $\text{Cov}(S) = I$, 对于 $Z(t) = W_0 AS(t) = \bar{A}S(t)$, 有

$$E\{ZZ^T\} = E\{\bar{A}SS^T \bar{A}^T\} = \bar{A} E\{SS^T\} \bar{A}^T = \bar{A} \bar{A}^T = I$$

即如果把上式的 $Z(t)$ 看成是新的观测信号, 则可以说: 白化使原来的混合矩阵 A 简化成一个新的正交矩阵 \bar{A} 。

3) ICA 判据: 在设计 ICA 算法的过程中, 最大的困难是如何可靠地验证源信号分量间的独立性, 为了度量这些分量的独立性, 需要给出独立性判据。下面给出 FastICA 的判据及计算流程。

FastICA 是一种快速 ICA (Fast ICA) 算法, 又称为固定点 (Fixed-Point) 算法, 是由芬兰赫尔辛基大学阿波·海韦里恩 (Aapo Hyvärinen) 和埃尔基·奥亚 (Erkki Oja) 等人提出来的, 该算法是基于定点递推算法得到的, 它对任何类型的数据都适用, 同时它的存在使得运用 ICA 分析高维的数据成为可能。FastICA 算法有基于四阶累积量、基于似然最大、基于负熵最大等形式。

前面已有类似的定义, 对于一个离散的随机变量 X , 其熵的定义为:

$$H(X) = - \sum P(X_i) \log(P(X_i)), i = 1, 2, \dots, n$$

其中, X_i 是 X 的可能取值, $P(X_i)$ 是 X 取不同值的概率。对于一个连续的随机变量 X , 其熵 (微分熵) 的定义为:

$$H(X) = - \int_a^b P(X) \log(P(X)) dX$$

我们可以利用熵的修正形式 (即负熵) 来度量非高斯性, 负熵定义如下:

$$N_g(Y) = H(Y_{\text{gauss}}) - H(Y)$$

其中, Y_{gauss} 是一与 Y 具有相同方差的高斯随机变量, $H(\cdot)$ 为随机变量的 (微分) 熵, 即根据信息理论, 在具有相同方差的随机变量中, 高斯分布的随机变量具有最大的微分熵, 当 Y 具有高斯分布时, $N_g(Y) = 0$; Y 的非高斯性越强, 其微分熵越小, 则对应的 $N_g(Y)$ 越大, 所以 $N_g(Y)$ 可以作为随机变量 Y 非高斯性的度量。

由于计算 $N_g(Y)$ 需要知道 Y 的概率密度分布函数, 这在实际中往往不可行, 因此采用如下近似公式, 即

$$N_g(Y) = \{E[g(Y)] - E[g(Y_{\text{gauss}})]\}^2$$

其中, $E[\cdot]$ 为均值运算符, $g(\cdot)$ 为非线性函数, 可取的非线性函数有 (包括但不限于):

□ $g(Y) = \tanh(aY)$, $1 \leq a \leq 2$, 一般取 $a = 1$ 。

□ $g(Y) = Ye^{-\frac{Y^2}{2}}$ 。

□ $g(Y) = Y^3$ 。

FastICA 实际上是一种寻找 $W^T Z (Y = W^T Z)$ 的非高斯最大的不动点迭代方案。为了推导近似牛顿法, 首先 $W^T Z$ 的近似负熵的极大值通常在 $E\{g(W^T Z)\}$ 的极值处取得, 根据拉格朗日条件, $E\{g(W^T Z)\}$ 在约束 $E\{(W^T Z)^2\} = \|W\|^2 = 1$ 条件下的极值, 是在那些使得拉格朗日乘子式的梯度为 0 的点处取得的:

$$E\{Zg(W^T Z)\} + \beta W = 0$$

其中, β 为拉格朗日乘子, 为利用牛顿迭代法求解, 令 $F = E\{Zg(W^T Z)\} + \beta W$, 则

$$\frac{\partial F}{\partial W} = E\{ZZ^T g'(W^T Z)\} + \beta I$$

为了简化计算过程, 对上式进行近似, 因为数据已经过球化处理, 因此:

$$E\{ZZ^T g'(W^T Z)\} \approx E\{ZZ^T\} E\{g'(W^T Z)\} = E\{g'(W^T Z)\} I$$

于是得到近似的牛顿迭代算法:

$$W \leftarrow W - \frac{E\{Zg(W^T Z)\} + \beta W}{E\{g'(W^T Z)\} + \beta I}$$

上式两边同时乘以 $E\{g'(W^T Z)\} + \beta$, 进一步简化为:

$$W \leftarrow E\{Zg(W^T Z)\} - E\{g'(W^T Z)\} W$$

每次迭代完, 就对 W 进行标准化, 以上就是 FastICA 算法中不动点迭代的基本过程。完整的 FastICA 算法流程描述如下:

输入: 观测数据 X

输出: W_F

Step1: 对观测数据 X 进行中心化, 使它的均值为 0。

Step2: 对数据进行白化, $X \rightarrow Z$ 。

Step3: 选择需要顾及的分量的个数 m , 设迭代次数 $p \leftarrow 1$ 。

Step4: 选择一个初始权向量 (随机) W_F 。

Step5: 令 $W_F = E\{Zg(W_F^T Z)\} - E\{g'(W_F^T Z)\} W_F$ 。

Step6: $W_F = W_F - \sum_{j=1}^{p-1} (W_F^T W_j) W_j$ 。

Step7: 如果 W_F 不收敛, 则返回 Step5。

Step8: $p \leftarrow p + 1$, 如果 $p \leq m$, 则返回 Step4。

需要注意的是, FastICA 算法具有如下几个典型特点 (包括但不限于):

1) 对观测信号去均值是 ICA 算法最基本和最必需的预处理步骤, 该预处理只是为了简化

ICA 算法,并不意味着均值不能被估计出来。

2) 一般情况下所获得的数据都具有相关性,通常都要求对数据进行初步的白化或球化处理(经过白化处理后,算法的收敛性要更好),因为白化处理可去除各观测信号之间的相关性,从而简化后续独立分量的提取过程。

3) 对多个独立分量的估计,需要将最大非高斯性的方法加以扩展。独立分量可被逐个估计出来,在探索性数据分析里是非常有用的,这类似于做投影追踪,在仅需要估计几个(不是全部)独立分量的情况下,能极大地降低计算量。

4. 基于数字信号处理的变换

前文介绍的 PCA、LDA、ICA 本质上都是基于一种线性变换以投影映射的方式将原始特征转变为维度比较低的特征(组合)。在线性变换的基础上又引申出很多非线性变换,比如 Kernel PCA、Kernel FDA 及 Manifold Learning(流形学习,找到流形上的低维坐标,利用流形学上的局部结构进行降维的方法,主要有 ISOMAP、LLE、Laplacian Eigenmap、LPP 等),此处不再赘述。本节将重点介绍用于特征提取的数字信号处理方法,主要包括傅里叶变换 DFT(即快速傅里叶变换 FFT)、离散余弦变换 DCT(对偶的一种变换是离散正弦变换 DST)、离散小波变换 DWT 等。上述所有的这些变换其目标都是时频变换,即将信号(原始特征)从时域表示方式转换为频域表示方式。

所谓信号是指一个或多个独立变量的函数(自变量可以是时间、距离、速度、位置、温度或压力等),该函数含有物理系统的信息以表示物理系统状态或行为。根据自变量是连续的还是离散的又可将信号分为连续时间信号或离散时间信号,其中自变量离散因变量也离散的称为数字信号,自变量连续因变量也连续的称为模拟信号。当然,也可以从其他维度对信号进行分类,比如随机信号、确定信号、一维信号、二维信号等,此处不再赘述。

数字信号一般表示为 $x(nT)$, 其中, T 表示采样间隔,即在第 n 个采样点的值是 $x(nT)$ 。由于 T 只是一个采样间隔,没有更多的物理意义和数学意义,因此更一般的表示是将 T 忽略,直接写成 $x(n)$ 。我们将所有表示成这种形式的(数字信号)都称为时域信号,自变量未必是狭义的时间,可以是任何一种独立变量,比如前文提到的距离、速度、位置、温度、压力等。

傅里叶变换的物理意义在于:对于任何一个周期信号,都可以展开成为若干个不同权重的频率不同的正弦波的累加和。这意味着任何一个诸如 $x(n)$ 的时域信号都可以展开成不同权重组成的一组向量,其中的每一个值均表示不同频率正弦波的分量,我们将这个转换过程称为时频变换,即将信号从时域表示(以广义的“时间”为自变量)转换为频域表示(以频率为自变量)。在实际操作中,频率在 $(0, 2\pi)$ 之间等间距进行采样,采样间隔为 $\frac{2\pi}{N}$, 其中 N 的大小取决于时域信号 $x(n)$ 的长度。

(1) DFT

对于给定的一维数字信号 $x(n)$, 其离散傅里叶变换 DFT 定义如下:

$$\text{DFT}[x(n)] = X(e^{-j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \quad k = 0, 1, 2, \dots, N-1$$

令 $W_N = e^{-j\frac{2\pi}{N}}$, 则上式可以简化为:

$$\text{DFT}[x(n)] = X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad k = 0, 1, 2, \dots, N-1$$

将上式写成矩阵表示形式, 可得到:

$$\begin{pmatrix} X(0) \\ X(1) \\ \vdots \\ X(N-1) \end{pmatrix} = \begin{pmatrix} W_N^{0 \times 0} & W_N^{0 \times 1} & \cdots & W_N^{0 \times (N-1)} \\ W_N^{1 \times 0} & W_N^{1 \times 1} & \cdots & W_N^{1 \times (N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ W_N^{(N-1) \times 0} & W_N^{(N-1) \times 1} & \cdots & W_N^{(N-1) \times (N-1)} \end{pmatrix} \begin{pmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{pmatrix}$$

DFT 变换是可逆的, 其对应的反变换是:

$$\text{DFT}[X(k)] = x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn}, \quad k = 0, 1, \dots, N-1$$

从上式可以看出, DFT 变换及反变换的计算涉及 N^2 次 W_N 的计算, 计算量很大, 阻碍了其在长信号处理中的应用, 由于 W_N 矩阵具有对称性和周期性, 可以充分利用此特点对上述计算过程进行简化, 包括按时间基抽取和按频率基抽取, 本文以按时间基抽取的思路进行一个推导思路的简单介绍, 该思路的基本出发点是: 基于较小的 DFT 计算一个较大的 DFT。即首先将长度为 N 的序列 $x(n)$ 划分为 M 个长度为 L 的较小序列, 然后做这 M 个较小序列的 L 点 DFT, 最后利用这 M 个较小的 L 点 DFT 组合成 N 点的 DFT。

将序列 $x(n)$ 划分为 M 个长度为 L 的较小序列, 即将原始一维向量转换为 $L \times M$ 的二维数组, 即:

$$\{x(0), x(1), \dots, x(N-1)\} \rightarrow \begin{pmatrix} x(0) & x(1) & \cdots & x(M-1) \\ x(M) & x(M+1) & \cdots & x(2M-1) \\ \vdots & \vdots & \ddots & \vdots \\ x(LM-M) & x(LM-M+1) & \cdots & x(LM-1) \end{pmatrix}$$

其中, 每一列均是长度为 L 的较小序列, 该序列中的每个元素均可表示为 $x(l, m)$, 其中 $0 \leq m < M$; $0 \leq l < L$, 且 $n = Ml + m$ 。相应的, DFT 的结果 $X(k)$ 也要变为二维的矩阵表示, 即:

$$\{X(0), X(1), \dots, X(N-1)\} \rightarrow \begin{pmatrix} X(0) & X(1) & \cdots & X(M-1) \\ X(M) & X(M+1) & \cdots & X(2M-1) \\ \vdots & \vdots & \ddots & \vdots \\ X(LM-M) & X(LM-M+1) & \cdots & X(LM-1) \end{pmatrix}$$

其中, 序列的每个元素均可表示为 $X(p, q)$, 其中 $0 \leq q < M$, $0 \leq p < L$, 特别注意: $k = Lq + p$

(先对 $x(l, m)$ 矩阵的每一列进行 DFT 操作, 然后再对每一行进行 DFT 操作)。则 $X(k) =$

$\sum_{n=0}^{N-1} x(n) W_N^{kn}$ 可转换为:

$$X(p, q) = \sum_{m=0}^{M-1} \sum_{l=0}^{L-1} x(l, m) W_N^{(Ml+m)(Lq+p)} = \sum_{m=0}^{M-1} \sum_{l=0}^{L-1} x(l, m) W_N^{(MLq+Mlp+mLq+mp)}$$

将 $N = ML$ 代入并考虑到 W_N 是以 N 为周期的, 因此上式可以简化为:

$$X(p, q) = \sum_{m=0}^{M-1} \sum_{l=0}^{L-1} x(l, m) W_N^{(Mlp+mLq+mp)} = \sum_{m=0}^{M-1} \{ W_N^{mp} \sum_{l=0}^{L-1} x(l, m) W_N^{Mlp} \} W_N^{mLq}$$

由于 $W_N^{Mlp} = W_L^{lp}$ 及 $W_N^{mLq} = W_M^{mq}$, 因此上式可进一步简化为:

$$X(p, q) = \sum_{m=0}^{M-1} \{ W_N^{mp} \sum_{l=0}^{L-1} x(l, m) W_L^{lp} \} W_M^{mq}$$

上式表明 $X(p, q)$ 的计算过程是:

1) 首先对各列计算 L 点的 DFT 数组(备注: $\sum_{l=0}^{L-1} x(l, m) W_L^{lp}$)。

2) 对数组乘以 W_N^{mp} 进行修正(备注: $W_N^{mp} \sum_{l=0}^{L-1} x(l, m) W_L^{lp}$)。

3) 求出每行 M 点的 DFT。

当 $N = R^V$ 时, 我们称为基 $-R$ 的快速傅里叶变换算法。令上面推导过程中的 $M = 2$, $L = \frac{N}{2}$, 即将原始序列拆分为 2 个 $\frac{N}{2}$ 点的序列 (即奇数部序列和偶数部序列), 即

$$x_e(n) = x(2n), x_o(n) = x(2n+1), 0 \leq n \leq \frac{N}{2} - 1$$

令偶数部序列的 DFT 是 $X_e(k)$, 令奇数部序列的 DFT 是 $X_o(k)$, 则上式可以简化为:

$$X(p, q) = W_N^{mp} \sum_{l=0}^{L-1} \{ x(l, m) W_L^{lp} \} W_M^{mq} \Big|_{m=0} + W_N^{mp} \sum_{l=0}^{L-1} \{ x(l, m) W_L^{lp} \} W_M^{mq} \Big|_{m=1}$$

进一步简化上式得到

$$X(p, q) = \sum_{l=0}^{L-1} x(l, m) W_L^{lp} + W_N^p W_M^q \sum_{l=0}^{L-1} x(l, m) W_L^{lp} = X_e(k) + W_N^k X_o(k)$$

上式的数学意义在于: 对于一个长度为 N 的序列而言, 其 DFT 可以由它的偶数部序列和奇数部序列组成, 即

$$X(k) = X_e(k) + W_N^k X_o(k)$$

$$X\left(k + \frac{N}{2}\right) = X_e(k) + W_N^{k+\frac{N}{2}} X_o(k)$$

对于上面的公式, 有一个名为蝶形计算单元的图示方法可以更好地描述其计算过程, 如图 7-8 所示。

这个过程可以一直重复, 在每一步中都对序列进行抽取, 并且较小的 DFT 会被合并, 当经过 $V(N = 2^V)$ 步以后就具有 N 个一点的序列, 此时 DFT 终止, 这个过程

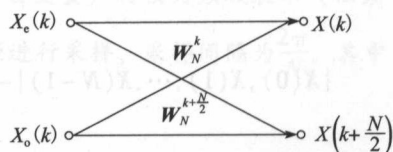


图 7-8 蝶形计算单元

就称为按时间抽取的（基-2）快速傅里叶变换（FFT）。

由于傅里叶变换具有物理上的直观性及数学上的完美性，特别是在计算方面有高效的算法支撑，因此傅里叶变换成为信号处理中广为流传和应用的算法之一。在进行特征提取的过程中，通过傅里叶变换获取信号的频域特征往往也能取得很好的效果。

值得注意的是，由于傅里叶变换是基于e变换核的，因此傅里叶变换的结果往往是复数，显然这在实际应用中是无法使用的，因此，一般基于傅里叶变换的特征提取，需要将该复数结果的幅度和相位分别提出，然后利用幅度或相位作为应用中的特征表示。

(2) DCT

对于一个一维信号 $f(x)$ ，其DCT变换为：

$$C(u) = a(u) \sum_{x=0}^{N-1} f(x) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \quad u = 0, 1, \dots, N-1$$

其中 $a(u) = \begin{cases} \sqrt{\frac{1}{N}}, & u=0 \\ \sqrt{\frac{2}{N}}, & u=1, 2, \dots, N-1 \end{cases}$ ，其对应的逆变换是：

$$f(x) = \sum_{u=0}^{N-1} a(u) C(u) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \quad x = 0, 1, \dots, N-1$$

对于一个二维信号，其DCT变换为：

$$C(u, v) = a(u)a(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right]$$

其对应的逆变换是：

$$f(x, y) = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} a(u)a(v) C(u, v) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right]$$

DCT实际上是DFT的实数部分，是一种可逆的实数变换，即实数信号到实数DCT系数之间的映射。DCT的典型特点在于：相比较于DFT，它有更强的信息集中能力，DCT用于声音和图像处理时，大部分能量都集中在极少数系数上，从而可以提高编码的效率。

图7-9显示了基于DCT变换的一组基础实验流程。

Step1：原始图像Image-1经过DCT变换后得到DCT变换矩阵Image-DCT1。

Step2：将Image-DCT1直接进行IDCT操作，得到Image-2，由于DCT变换是可逆的，因此Image-2和Image-1应该是完全一致的。

Step3：如果将Image-DCT1的下三角矩阵全部置0可得到Image-DCT2，然后再将Image-DCT2进行IDCT变换，可得到Image-3。

有证据表明：即便Image-DCT2仅保留有限的上三角矩阵（比如仅占全部 $M \times N$ 元素的15%），IDCT的结果Image-3与原始图像Image-1在肉眼上也无法区别。这是由于DCT变换具有的信息集中能力强这个特点所导致的，原始信号的大部分信息集中在变换矩阵的前几个上三角矩阵元素中（对于一维信号，集中在一维信号的前几位）。也正是由于这个原因，DCT常

被用于图像压缩、数字水印、信息隐藏等很多领域。从特征提取的角度而言,能够从原始信号中提取出极少量的可表征原始信号的特征,也是大有裨益的,这也是 DCT 常用于特征表示和特征提取的重要原因。

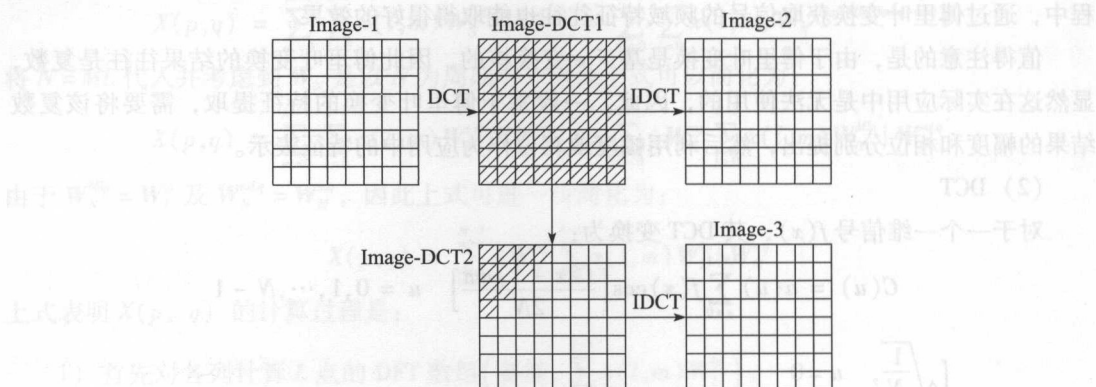


图 7-9 DCT 基础实验

以图像匹配为例,一般情况下,在两幅图像中,很难准确找到两者之间对应的点。实际处理过程中,不同时间、地点、环境下得到的图像也不同,例如图像的转置、镜像、亮度或对比度的变化,会导致图像匹配的难度增加。二维离散余弦变换的变换矩阵有以下特性:

- ①变换后矩阵的能量主要集中在其左上角,为图像的直流成分和低频成分,并且矩阵的各元素是不相关的。
- ②离散余弦变换可逆,变换后的图像矩阵可通过逆 DCT 变换得到恢复,使得大部分图像信息可由矩阵左上角的部分数据恢复得到。
- ③矩阵稳定且具有健壮性,图像受到较小的干扰,或者环境的亮度和对比度在某一范围内做线性变化时,矩阵的元素特别是左上角的元素不会发生大的变化。
- ④图像在非仿射变换如镜像变换和转置变换时,矩阵会发生转置或矩阵中的元素符号会变化,但变换后矩阵的绝对值不变。所有这些特点都使得 DCT 在类似场景下的特征提取和特征表示方面具有天然的优势。

(3) DWT

前述的 FFT 和 DCT 本质上都是一种时间亚元的时频变换,即:信号从时域变换到频域后,变换后的信号已经没有了时间(域)的下标,这在很多场合都会受到限制。针对上述变换的此方面缺陷,小波变换被提到日程中来。

短时傅里叶变换(Short Time Fourier Transform, STFT)是为了克服传统傅里叶变换时间亚元的缺陷而进行的一种改进尝试,其基本思想可简单描述为:用一个时频局部化的窗口函数,将原始信号 $X(t)$ 划分为若干多个短时间间隔内的平稳(伪平稳)信号,然后对每一片段的信号进行傅里叶变换,即原始信号 $X(t)$ 被转换为

$$X'(t) = X(t)G(t - \tau)$$

其中, $G(t - \tau)$ 是以时间 τ 为中心的窗函数。短时傅里叶变换使用的这个窗函数是一个固定

的窗函数,但是窗函数一旦确定了以后,短时傅里叶变换的分辨率也就确定了。如果要改变分辨率,则需要重新选择窗函数。短时傅里叶变换的一个典型缺陷是不能同时兼顾频率与时间分辨率的需求,比如当信号变化剧烈时,要求窗函数有较高的时间分辨率(窄);而波形变化比较平缓时,则要求窗函数有较高的频率分辨率(宽),短时傅里叶变换窗函数的时间与频率分辨率不能同时达到最优。小波变换之所以得到工业界和学术界的普遍关注,其重要原因在于:通过对原始信号进行多轮次的小波变换可以达到在时间与频率上的同时最优。

前面介绍的傅里叶变换及离散余弦变换本质上都是将原始信号在一组正交基上展开,小波变换也不例外,所谓小波变换就是:

1) 将原始信号在一组小波基上展开得到一组小波系数,该系数的意义在于:

①通过此系数结合当前的小波基可以无失真地恢复出原始信号,这意味着小波变换是完全可逆的,也就是说:在实际应用中,此组系数可以作为表征原始信号的一组特征来使用。

②该系数在逻辑上被等分为对应不同物理意义的两部分,一部分系数反映原始信号的低频成分,另外一部分系数反映原始信号的高频成分,也就是说:在实际应用中,可以根据需求有目的地选取低频系数、高频系数或全部选择。

2) 考察第1步提取的高频系数和低频系数,如果不满足实际应用场景需求,则将上一组小波基进行变换后(变换规则后议)形成新的一组小波基,然后重复第1步。

因此,一个完整的小波变换可以用图7-10表示。如图7-10所示,提到小波变换,就意味着需要进行 N 次变换(至于 N 是多少,需要根据具体的应用需求来设置),每次变换就将原始信号分为低频和高频两个部分(事实上是对应低频和高频的两组小波系数),根据实际应用需求对原先的低频或高频继续进行分解(如果每次只针对低频部分进

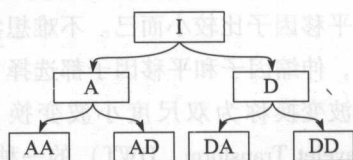


图7-10 小波变换示意图

行继续分解,这样的小波分解方式称为小波分解树;如果每次对低频系数和高频系数均进行分解,则称为小波包分解树)。在实际应用中,可根据实际场景下的数据特点(往往通过对原始数据进行多组实验)来确定小波变换的次数 N ,以及使用哪些小波系数作为原始数据的特征表示。

为了达到在时频域同时定位的目标,小波变换对小波基函数及小波基函数的变换有着特殊的需求,具体而言:

1) 小波基函数是在小波母函数的基础上生成的,而能够称得上是小波母函数(也称为基本小波)的函数应当满足紧支性条件和容许性条件:

令 $\varphi(t)$ 为一平方可积函数(容许性条件),即 $\varphi(t) \in L^2(R)$,如果其傅里叶变换 $\Psi(w)$ 满足紧支性条件,即

$$\int_R \frac{|\Psi(w)|^2}{w} dw < \infty$$

则称 $\varphi(t)$ 为小波母函数, 小波母函数的波形可以是不规则的, 也可以是规则的 (比如对称的), 但是小波母函数在有限的持续时间上具有突变的频率和振幅, 在整个时间范围里的幅度平均值为零 (证明从略)。

2) 将小波母函数 $\varphi(t)$ 进行伸缩和平移就可得到一组小波基函数 (每轮次的小波变换, 事实上就是通过对小波母函数的伸缩和平移得到的), 令伸缩因子 (也称为尺度因子) 为 a , 平移因子为 τ , 则依赖参数 a 和 τ 的小波基函数可定义为:

$$\varphi_{a,\tau}(t) = \frac{1}{\sqrt{a}} \varphi\left(\frac{t-\tau}{a}\right)$$

其中, $a > 0$ 且 $\tau \in R$ 。小波的缩放因子 a 与信号频率之间的关系可以这样来理解: 缩放因子小, 则表示小波比较窄, 度量的是信号细节; 相反, 缩放因子大, 则表示小波比较宽, 度量的是信号的形似。

基于上述小波基, 将任意 $L^2(R)$ 空间的信号 $x(t)$ 在小波基下进行展开称为 $x(t)$ 的连续小波变换 (CWT), 定义为:

$$W(x(t), \varphi_{a,\tau}(t)) = \frac{1}{\sqrt{a}} \int_R x(t) \overline{\left(\frac{t-\tau}{a}\right)} dt$$

其中 a 是伸缩因子, τ 是平移因子, 并且都是连续的。

在计算连续小波变换时, 实际上也是用离散的数据来进行计算的, 只是所用的伸缩因子和平移因子比较小而已。不难想象, 连续小波变换的计算量是惊人的。为了解决计算量的问题, 伸缩因子和平移因子都选择 2^j ($j > 0, j \in Z$) 的倍数。使用这样的伸缩因子和平移因子的小波变换称为双尺度小波变换 (Dyadic Wavelet Transform), 它是离散小波变换 (Discrete Wavelet Transform, DWT) 的一种形式。

任意 $L^2(R)$ 空间的信号 $x(t)$ 的离散小波变换 (DWT) 定义为:

$$W_{x(t)}(j, k) = \int_R x(t) \overline{\varphi_{j,k}(t)} dt$$

其中, $\varphi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \varphi\left(\frac{t}{2^j} - k\right)$, 注意: 这里的离散化是针对伸缩因子和参数因子而言的。

下面以哈尔小波变换 (最简单的小波变换) 来解释小波变换的过程, 哈尔小波变换是基于哈尔母小波的变换, 哈尔母小波是基于哈尔基函数的一种变形, 哈尔基函数是 1990 年提出的, 由一组分段常值函数组成的函数集, 该函数集定义在 $[0, 1)$ 范围内, 分段常值在一定的范围内是 1, 其他范围内为 0, 定义为:

$$\varphi(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{其他} \end{cases}$$

为了表示不同向量空间的向量, 哈尔基函数的尺度函数表示为:

$$\varphi_i^j(x) = \varphi(2^j x - i)$$

其中, i 为平移参数, j 是尺度因子, 使用 $\varphi_i^j(x)$ 表示的向量空间表示为 V^j , 表 7-2 给出了 j

的不同取值可以表示的向量空间及对应的哈尔基函数。

表 7-2 向量空间 V^j 与对应的哈尔基函数

j	向量空间	哈尔基函数
0	V^0	$\varphi_0^0(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{其他} \end{cases}$
1	V^1	$\varphi_0^1(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ 0, & \text{其他} \end{cases}, \quad \varphi_1^1(x) = \begin{cases} 1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{其他} \end{cases}$
2	V^2	$\varphi_0^2(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{4} \\ 0, & \text{其他} \end{cases}, \quad \varphi_1^2(x) = \begin{cases} 1, & \frac{1}{4} \leq x < \frac{1}{2} \\ 0, & \text{其他} \end{cases},$ $\varphi_2^2(x) = \begin{cases} 1, & \frac{1}{2} \leq x < \frac{3}{4} \\ 0, & \text{其他} \end{cases}, \quad \varphi_3^2(x) = \begin{cases} 1, & \frac{3}{4} \leq x < 1 \\ 0, & \text{其他} \end{cases}$
...

由表 7-2 其实可以清晰地看到：由于这些向量都是在 $[0, 1)$ 上定义的，因此在 V^j 向量空间中的每个向量都被包含在 V^{j+1} 中，即

$$V^0 \subset V^1 \subset V^2 \subset \dots \subset V^j \subset V^{j+1}$$

基于哈尔基函数进行变形得到的哈尔小波函数用 $\psi(x)$ 表示，定义如下：

$$\psi(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{其他} \end{cases}$$

为了表示不同向量空间的向量，哈尔小波函数的尺度函数 $\psi_i^j(x)$ 表示为：

$$\psi_i^j(x) = \psi(2^j x - i)$$

其中， i 为平移参数， j 是尺度因子，使用 $\psi_i^j(x)$ 表示的向量空间为 W^j 。表 7-3 给出了 j 的不同取值可以表示的向量空间及对应的哈尔小波基函数。

表 7-3 向量空间 W^j 与对应的哈尔小波基函数

j	向量空间	哈尔小波基函数
0	W^0	$\psi_0^0(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{其他} \end{cases}$
1	W^1	$\psi_0^1(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{4} \\ -1, & \frac{1}{4} \leq x < \frac{1}{2} \\ 0, & \text{其他} \end{cases}, \quad \psi_1^1(x) = \begin{cases} 1, & \frac{1}{2} \leq x < \frac{3}{4} \\ -1, & \frac{3}{4} \leq x < 1 \\ 0, & \text{其他} \end{cases}$

(续)

j	向量空间	哈尔小波基函数
2	W^2	$\psi_0^2(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{8} \\ -1, & \frac{1}{8} \leq x < \frac{1}{4} \\ 0, & \text{其他} \end{cases}$ $\psi_1^2(x) = \begin{cases} 1, & \frac{1}{4} \leq x < \frac{3}{8} \\ -1, & \frac{3}{8} \leq x < \frac{1}{2} \\ 0, & \text{其他} \end{cases}$ $\psi_2^2(x) = \begin{cases} 1, & \frac{1}{2} \leq x < \frac{5}{8} \\ -1, & \frac{5}{8} \leq x < \frac{3}{4} \\ 0, & \text{其他} \end{cases}$ $\psi_3^2(x) = \begin{cases} 1, & \frac{3}{4} \leq x < \frac{7}{8} \\ -1, & \frac{7}{8} \leq x < 1 \\ 0, & \text{其他} \end{cases}$
...

假设两个信号的数值分别是 a 和 b , 则它们的和与差分别是 $s = a + b$ 和 $d = a - b$, 显然, 从 s 和 d 很容易恢复出原始数值 a 和 b , 其中: $a = \frac{s+d}{2}$, $b = \frac{s-d}{2}$ 。

现在的问题是: 假设有一幅分辨率只有 4 个像素的一维图像, 对应的数据是 $[9, 7, 3, 5]$, 计算它的哈尔小波变换 (系数)。步骤如下:

Step1: 求均值 (Averaging)

计算相邻像素对的平均值, 得到一幅分辨率比较低的新图像, 它的像素数目变成了 2 个, 即新图像的分辨率是原来的 $\frac{1}{2}$, 相应的像素值为 $[8, 4]$ 。

Step2: 求差值 (Differencing)

计算每个像素值与 Step1 计算的平均值的差值 (事实上, 对于相邻的 2 个像素, 只需要计算其中的一个像素, 比如第一个像素值和平均值的差值即可), 得到 $[1, -1]$ 。用 Step1 的计算结果表示原始图像 (数据) 时, 原始数据的信息已经丢失, 这意味着无法利用 Step1 的计算结果恢复出原始数据, 而如果把 Step1 和 Step2 的结果联合在一起得到的 $[8, 4, 1, -1]$ 就能够不失真地恢复出原始数据, 其中 Step1 得到的结果反映的是原始信号的低频, Step2 得到的结果是原始信号的高频 (细节部分)。

Step3: 重复 Step1 和 Step2

把由第 1 步分解得到的图像进一步分解成分辨率更低的图像和细节系数。在这个例子中, 分解到最后, 就是用 $[6, 2, 1, -1]$ 表示原始的图像数据, 其中 6 是均值, 而另外三个值均是不同层次分解的细节系数。

上述的分解过程就是典型的哈尔小波分解过程 (其他小波基的分解也采用类似的步骤)。图 7-11 给出了分解示意图。

求均值和差值的过程实际上就是一维小波变换的过程, 现在用数学的方法重新描述小波变换的过程:

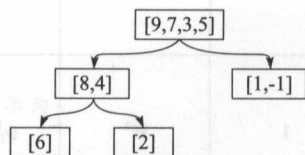


图 7-11 哈尔小波分解示意图

1) 用 V^2 中的哈尔基表示, 原始信号 $[9, 7, 3, 5]$ 可以表示为:

$$I = c_0^2 \varphi_0^2(x) + c_1^2 \varphi_1^2(x) + c_2^2 \varphi_2^2(x) + c_3^2 \varphi_3^2(x) = 9\varphi_0^2(x) + 7\varphi_1^2(x) + 3\varphi_2^2(x) + 5\varphi_3^2(x)$$

其中的系数 c_0^2 、 c_1^2 、 c_2^2 、 c_3^2 是 4 个正交的像素值, 更直观的图像表示如图 7-12 所示。

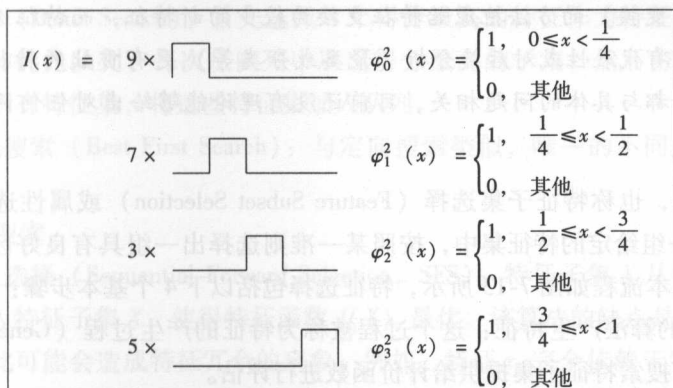


图 7-12 V^2 表示示意图

2) 用 V^1 中的哈尔基和 W^1 中的哈尔小波基表示第一层小波变换的结果, 用 V^1 表示 $[8, 4]$, W^1 表示 $[1, -1]$, 则

$$I = c_0^1 \varphi_0^1(x) + c_1^1 \varphi_1^1(x) + d_0^1 \psi_0^1(x) + d_1^1 \psi_1^1(x) = 8\varphi_0^1(x) + 4\varphi_1^1(x) + \psi_0^1(x) - \psi_1^1(x)$$

图示略。

3) 用 V^0 中的哈尔基、 W^0 中的哈尔小波基、 W^1 中的哈尔小波基表示, 用 V^0 表示 $[6]$, W^0 表示 $[2]$, W^1 表示 $[1, -1]$, 则

$$I = c_0^0 \varphi_0^0(x) + d_0^0 \psi_0^0(x) + d_0^1 \psi_0^1(x) + d_1^1 \psi_1^1(x) = 6\varphi_0^0(x) + 2\psi_0^0(x) + \psi_0^1(x) - \psi_1^1(x)$$

图示略。

在具体的实践过程中, 必须根据实际应用场景的特点进行小波母函数的选择和小波变换次数的选择。对于小波母函数的选择, 由于可以选择的小波母函数有很多, 因此必须根据待分析数据的特点选择有针对性的小波母函数 (也可以利用后文介绍的特征选择策略指导小波母函数的选择); 针对后者, 也必须根据实际情况分析需求, 进行实验, 理性地选择小波变换的次数 (可以利用下文介绍的特征选择策略进行选择)。

7.4.3 特征选择

1. 特征选择过程

在机器学习的实际应用中, 特征数量往往较多 (特征向量维度非常大), 其中可能存在与应用目标不 (太) 相关的特征, 或者特征之间存在相互依赖的现象, 容易导致诸如训练时间长、模型过于复杂、模型的泛化能力弱等问题。因此在进行机器学习与数据挖掘之前有必要进行特征选择 (或者属性选择)。

在实际应用中,特征提取和特征选择经常联合使用,两者都是从原始特征中找出最有效(同类样本的不变性、不同样本的鉴别性、对噪声干扰的鲁棒性)的特征,从而达到降低维度、提取有效信息、压缩特征空间、减少计算量、发现更有意义的潜在变量等作用。特征提取专注于用映射(变换)的方法把原始特征变换为较少的新特征,而特征选择专注于从原始特征中挑选一些最有代表性或对后续分析(聚类或分类等)更有贡献的特征。通常而言,特征提取和特征选择都与具体的问题相关,目前还没有理论能够给出对任何问题都有效的特征提取与选择方法。

所谓特征选择,也称特征子集选择(Feature Subset Selection)或属性选择(Attribute Selection),是指从一组给定的特征集中,按照某一准则选择出一组具有良好区分特性的特征子集。特征选择的基本流程如图 7-13 所示,特征选择包括以下 4 个基本步骤:

- 1) 利用具体的算法产生特征:这个过程被称为特征的产生过程(Generation Procedure),负责从原始特征中搜索特征子集提供给评价函数进行评估。
- 2) 利用评价函数(Evaluation Function)评估特征子集的好坏,评价函数是评价一个特征子集好坏程度的一个准则。
- 3) 通过设置的停止准则(Stopping Criterion)决定是否要继续搜索特征子空间,停止准则与评价函数相关的,一般是一个阈值,当评价函数值达到这个阈值后就可停止搜索。
- 4) 验证过程(Validation Procedure):在验证数据集上验证选择出的特征子集在目标应用方面的有效性。

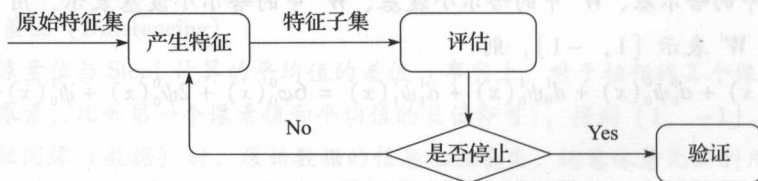


图 7-13 特征选择流程

2. 产生过程

特征产生负责从原始的所有特征中搜索特征子集提供给评价函数进行评估,一般使用基于搜索的方法进行特征选择,搜索的算法分为完全搜索(Complete Search)、启发式搜索(Heuristic Search)、随机搜索(Random Search)3 大类,以下是简单介绍。

(1) 完全搜索

完全搜索分为穷举搜索(Exhaustive Search)与非穷举搜索(Non-Exhaustive Search)两类,具体而言有 4 类:

- 1) 广度优先搜索(Breadth First Search):指的是随机从一个特征出发,广度优先遍历特征子空间,属于穷举搜索,每一个特征都有两种可能性,被选中 and 没有被选中,因而这种方法的时间复杂度是 $O(2^n)$,虽然可以找出最优解,但速度太慢,实用性不高。

2) 分支限界搜索 (Branch and Bound): 在穷举搜索的基础上加入分支限界, 缩小需要搜索的空间。例如: 若断定某些分支不可能搜索出比当前找到的最优解更优的解, 则可以剪掉这些分支。

3) 定向搜索 (Beam Search): 首先选择 N 个得分最高的特征作为特征子集, 将其加入一个限制最大长度的优先队列, 每次从队列中取出得分最高的子集, 然后穷举向该子集加入 1 个特征后产生的所有特征集, 将这些特征集加入队列。

4) 最优优先搜索 (Best First Search): 与定向搜索类似, 唯一的不同点是不限制优先队列的长度。

(2) 启发式搜索

1) 序列前向选择 (Sequential Forward Selection, SFS): 特征子集 X 从空集开始, 每次选择一个特征 x 加入特征子集 X , 使得特征函数 $J(X)$ 最优。该算法的缺点是只能加入特征而不能去除特征, 因此可能会造成特征冗余的现象。例如, 特征 x_i 完全依赖于特征 x_j 与 x_k , 可以认为如果加入了特征 x_j 与 x_k , 则 x_i 就是多余的。假设序列前向选择算法首先将 x_i 加入特征集, 然后又加入特征 x_j 与 x_k , 那么特征子集中就包含了多余的特征 x_i 。

2) 序列后向选择 (Sequential Backward Selection, SBS): 序列后向选择与序列前向选择正好相反。该方法从特征全集 O 开始, 每次从特征集 O 中剔除一个特征 x , 使得剔除特征 x 后评价函数值达到最优。它的缺点是特征只能去除不能加入。另外, SFS 与 SBS 都属于贪心算法, 容易陷入局部最优值。

3) 双向搜索 (Bidirectional Search): 使用序列前向选择从空集开始搜索, 同时使用序列后向选择从全集开始搜索, 当两者搜索到一个相同的特征子集 C 时即停止搜索。

4) 增 L 去 R 选择算法 (Plus- L Minus- R Selection): 该算法结合了序列前向选择与序列后向选择的思想, 有两种形式。第一种是算法从空集开始, 每轮先加入 L 个特征, 然后从中去除 $R(R < L)$ 个特征, 使得评价函数值最优; 第二种是算法从全集开始, 每轮先去除 R 个特征, 然后加入 $L(L < R)$ 个特征, 使得评价函数值最优。显然, 在此算法中, L 与 R 的选择是算法的关键。

5) 序列浮动选择 (Sequential Floating Selection): 序列浮动选择由增 L 去 R 选择算法发展而来, 不同之处在于序列浮动选择的 L 与 R 不是固定的, 而是变化的 (即所谓的“浮动”)。根据搜索方向的不同, 序列浮动选择算法有两种。第一种称为序列浮动前向选择 (Sequential Floating Forward Selection, SFFS), 该算法从空集开始, 每轮在未选择的特征中选择一个子集 X , 使加入子集 X 后评价函数达到最优, 然后在已选择的特征中选择一个子集 Z , 使剔除子集 Z 后评价函数达到最优; 第二种称为序列浮动后向选择 (Sequential Floating Backward Selection, SFBS), 与 SFFS 类似, 不同之处在于 SFBS 是从全集开始, 每轮先剔除特征, 然后再加入特征。

(3) 随机搜索

1) 随机产生序列算法 (Random Generation Plus Sequential Selection): 随机产生一个特征

子集,然后在该子集上执行 SFS 与 SBS 算法。可作为 SFS 与 SBS 的补充,用于跳出局部最优值。

2) 模拟退火算法 (Simulated Annealing): 模拟退火算法是爬山法的一个变形,当达到某一局部最优解时,以一定的概率跳转到某一次优解处,因此模拟退火算法一定程度上克服了序列搜索算法容易陷入局部最优值的缺点,但是若最优解的区域太小,则模拟退火难以求解。

3) 遗传算法 (Genetic Algorithm): 首先随机产生一批特征子集,并用评价函数给这些特征子集评分,然后通过交叉、突变等操作繁殖出下一代特征子集,并且评分越高的特征子集被选中参加繁殖的概率就越高。这样经过 N 代的繁殖和优胜劣汰后,种群中就可能产生了评价函数值最高的特征子集。

3. 评价函数

评价函数的作用是评价产生过程所提供的特征子集的好坏。根据其工作原理,评价函数主要分为筛选器 (Filter) 和封装器 (Wrapper) 两大类。

1) 筛选器通过分析特征子集内部的特点来衡量其好坏。筛选器一般用作预处理,与分类器的选择无关,常用的度量方法有相关性、距离、信息增益、一致性等。

运用相关性来度量特征子集的好坏是基于这样的假设,即好的特征子集所包含的特征应该与分类的相关度较高 (相关度高),而特征之间的相关度较低;运用距离度量进行特征选择是基于这样的假设,即好的特征子集应该使得属于同一类的样本距离尽可能地小,属于不同类的样本之间的距离尽可能地远;使用信息增益作为度量函数的动机在于假设存在特征子集 A 和特征子集 B ,分类变量为 C ,若 A 的信息增益比 B 大,则认为选用特征子集 A 的分类结果比 B 好,因此倾向于选用特征子集 A ;一致性指的是若样本 1 与样本 2 属于不同的分类,但在特征 A 和 B 上的取值完全一样,那么特征子集 $\{A, B\}$ 不应该选作最终的特征集。

筛选器由于与具体的分类算法无关,因此其在不同的分类算法之间的推广能力较强,而且计算量也较小。

2) 封装器实质上是一个分类器,封装器用选取的特征子集对样本集进行分类,分类的精度作为衡量特征子集好坏的标准。由于封装器在评价的过程中应用了具体的分类算法进行分类,因此其推广到其他分类算法的效果可能较差,而且计算量也较大。使用特定的分类器,用给定的特征子集对样本集进行分类,用分类的精度来衡量特征子集的好坏。

7.5 应用提示

度量方法的主要目标是让描述不同对象的一组数据具有可比性,实际应用下的度量方法选型往往需要兼顾数据的特征和后续分析目标的需求。以复杂网络中两个节点的相似性度量为例,杰卡德相似系数和欧氏距离均是可以使用度量方法,前者专注于两个节点在关系层面的紧密度,两个节点的共同邻居越多,表明两个节点在关系层面越接近。极端情况下,

两个节点的共同邻居与两个节点的邻居的全集完全一样,就表明这两个节点在关系属性方面完全相同。后者关注于每个节点的属性相似性,如果用特征向量来表示每个节点的属性,那么两个特征向量之间的距离就可以度量两个节点属性层面的相似性。

数据规范的主要目标是让描述不同对象的一组数据在进行比较(或参与计算)时是在一个相对公平的框架下进行的,出发点往往是集中在两个不同的粒度下进行的,相关策略包括(但不限于):

1) 对于一个特征向量 $\mathbf{X}(x_1, x_2, \dots, x_n)$ 中的每一个特征 x_i , 根据其取值范围对 x_i 进行最小—最大规范化。这种解决方案往往是基于对每一个特征 x_i 的物理语义及其取值范围的理解和明晰的基础,与应用是耦合的。

2) 对于一个特征向量 $\mathbf{X}(x_1, x_2, \dots, x_n)$, 可以根据每一个 x_i 的取值分布,对 x_i 进行最小—最大规范化、Z 分数规范化或按小数定标规范化。这种解决方案往往是基于特征向量中的每一个元素 x_i 在未来的计算(度量)中的权重相同的假设。

3) 实际上,上述两个步骤往往是耦合在一起进行的,先在每一个元素意义上进行规范,然后在整个数据上进行规范,为了兼顾每个元素所在的权重的不同,可以考虑在规范的过程中为每一个元素乘上相应的权重,这往往与应用场景直接相关。

从原始数据中提取出对后续数据分析有贡献的特征是数据建模及知识发现的重要基础,也是数据分析的第一步。特征提取的好坏直接影响后续的分析效果,这里所说的好或坏的评价依据一般有如下几个(包括但不限于):

1) 是否便于从原始数据中提取。应该注意到,并不是所有希望提取的特征在原始数据中都能够或很容易被提取出的,这与数据源质量、实时性要求直接相关。

2) 是否有助于后续的分析。比如此特征是否具有特异性,是否对后续的分析更有利;或者此特征是否更加精简,使得后续的分析在达到目标效果的同时,计算负载相对更低。

3) 提取的特征是否易于理解。用户天生而有的控制欲使然,往往希望提取的特征能够在物理上有一种语义解释,特征的可理解性有助于提高用户对整个系统的接纳程度。

显然,上述这3个指标要求的优先级是逐步降低的,这意味着在系统运行宿主环境及计算环境允许的情况下,要优先保证特征的特异性以便为后续的分析服务,在此基础上兼顾特征的可理解性指标。

特征对于数据分析的意义尤其重大,用任何表示重要的形容词来形容都不为过,甚至独立出专门的一个研究方向或工作岗位(角色)用于响应数据分析对特征的需求,即特征工程。既然称之为工程,其本质的定位在于:特征(提取)这件事情是一个工程化的事情,涉及多个环节,必须用工程化的手段管理这些逻辑串行(而实际上往往有交集甚至并行)的各项事务。特征工程涉及的主要环节包括:特征表示、特征提取与特征选择。

1) 特征表示的主要目的是从多个角度去研究从原始数据中能够提取出哪些有助于后续分类的特征来,往往需要从如下几个角度来展开工作(包括但不限于):

①对于任何一种数据类型,都应该有针对这些数据类型的普适性(常识性)特征可以提

取,这里的“普适”不是说这些特征对后续的分析一定有用,而是说它是一个潜在的、可供后续分析使用的候选特征集,比如:针对图像而言,形状、纹理、颜色是其最基本的特征,这意味着拿到一幅图像,下意识里的特征提取就应该是这些特征。

②鉴于任何一个原始数据描述的都是某一个领域的实体,而且后续数据分析也是围绕该领域的目标需求展开的。因此,充分与领域专家和用户交流和讨论,基于领域的知识进行特征的提取显然是非常有必要的。更重要的一点是,基于领域专家的知识(用户经验)提取的特征往往具有该领域认可 and 理解的物理语义,这对于最终用户理解和接纳目标系统大有裨益。

③在既没有常识经验也没有领域知识的支撑下,原始数据就是特征,这在原始数据维数非常小的情况下往往非常有效,即便是原始数据的维度很大,也可以利用后续的特征提取和特征选择手段,从中提炼和遴选出有贡献的特征,甚至通过这些无任何偏好指向提取和选择出来的特征更能够成为此类数据的有用经验,在未来加以复用。

2) 特征提取和特征选择的基本目标都是降低特征维度以增强后续的分析效果和性能,往往都是在从原始数据中解析出一些特征(基于特征表示)之后再进行的,两者的关系可以大致理解为(包括但不限于):

①基于数学意义上的转换(坐标投影)或变换(通过时频变换方法)等将原始的特征维度大幅降低,经过这种意义上的转换和变换以后,转换后的结果与原先的特征往往大相径庭;而特征选择是利用一些评估指标(可以是耦合后续的分析算法,也可以是独立的),从中挑选出对后续分析有贡献的特征来,经过这种意义上的选择之后,转换后的结果往往是原先特征的一个子集。

②特征提取和特征选择两种手段往往耦合在一起使用,比如先进行特征提取,然后进行特征选择,或者先进行特征选择,然后进行特征提取(注意:采用后者策略的时候,往往在评估指标的选择方面,应该采用筛选器,以使这种评估函数与后续的分析手段无关),特别值得一提的是,用于特征提取的时频变换方法,往往也可以直接应用于原始数据,比如直接对原始数据进行傅里叶变换或小波变换,然后将变换系数作为特征表示。

③由于特征提取是基于投影映射或变换的,因此变换后的结果往往在物理语义的理解上要远小于特征选择的结果。

3) 特别值得一提的是,虽然在逻辑上我们可以将特征表示、特征提取和特征选择理解为是串行执行的流程,但事实上,这三者是彼此耦合的。因此作为一名有经验的数据分析师,应该对数据、特征、分析手段及应用场景都具有极高的敏感度,才有可能从原始数据中遴选出对后续分析有价值的特征。

特别需要指出的是:特征的价值不仅局限于为后续的分析服务,特征本身或许就有很高的价值,比如:对于没有任何领域知识和先验经验的原始数据,将原始数据作为初始特征集,然后利用特征选择方法进行属性降维,最终的降维结果可以为未来类似数据的特征表示提供可借鉴的经验,甚至可以为进一步理解和分析其中的逻辑关系提供数据支撑。

另一方面,通过从原始数据中提炼出更加集约的特征,然后利用此特征进行数据建模,

从而达到揭示隐含在数据背后的规律和知识的目的,这是我们关注(和研究)特征的根本原因。但需要指出的是:无论是为了特征的提取和选择,还是仅仅获得数据本身,高效可信的数据组织及并发、高速的数据存取访问一般都是后续高效能计算的重要基础,而这一点往往会因为过于重视“特征→建模→知识”流程反而被(自然)忽略。

7.6 本章小结

度量方法选型、数据规范表示、特征工程是进行有效数据分析前必须要面对的三大问题,度量方法的目标是让普通的数据能够进行类似于标量一样的数学操作(当然,对于以向量出现的数据或特征而言,基本操作就是相似性度量);数据规范表示的目标是让不同数据源的数据平等、公平地在同一个计算逻辑中出现;特征工程的目标是通过行之有效的工程化方法和相应的技术手段从原始数据中提炼出有助于后续分析目标的特征。

本文顺序地介绍了上述三个问题,并对实际工作中各个环节可能的技术选型进行了扼要的介绍。需要注意的是,在实际应用中可以作为技术选型候选的度量方法、数据规范化方法、特征表示方法、特征提取方法、特征选择方法有很多,本章仅抛砖引玉地给出一些基本方法的介绍和应用提示,在实际应用中,需要结合具体应用的需求,进行有针对性的技术选型。

宋代禅宗大师青原行思提出参禅的三重境界是:第一重境界是“看山是山,看水是水”;第二重境界是“看山不是山,看水不是水”;第三重境界是“看山还是山,看水还是水”。

这三种参禅境界也反映出(被认为是)人生的三种境界,三种境界的演化乃内心(分析算法,比如分类或聚类,后文详议)使然,也是感觉和知觉(特征表示、提取与选择)使然。

本章参考文献

- [1] Abdi H, Williams L J. Principal Component Analysis [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459.
- [2] Cayton L. Algorithms for Manifold Learning [J]. Univ. of California at San Diego Tech. Rep, 2005: 1-17.
- [3] Cooley J W, Lewis P A W, Welch P D. The Fast Fourier Transform and Its Applications [J]. Education, IEEE Transactions on, 1969, 12(1): 27-34.
- [4] Dash M, Liu H. Feature Selection for Classification [J]. Intelligent Data Analysis, 1997, 1(1): 131-156.
- [5] Herault J, Jutten C. Space or Time Adaptive Signal Processing by Neural Network Models [C]. Neural networks for computing. AIP Publishing, 1986, 151(1): 206-211.
- [6] Hyvärinen A, Oja E. Independent Component Analysis: Algorithms and Applications [J]. Neural networks, 2000, 13(4): 411-430.

- [7] Izenman A J. Modern Multivariate Statistical Techniques [M]. New York: Springer, 2008.
- [8] Jain A, Zongker D. Feature Selection: Evaluation, Application, and Small Sample Performance [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997, 19(2): 153-158.
- [9] Kira K, Rendell L A. The Feature Selection Problem: Traditional Methods and A New Algorithm [C]. Proceedings of The Tenth National Conference on Artificial Intelligence, 1992: 129-134.
- [10] Liu Y, Zheng Y F. FS_SFS: A Novel Feature Selection Method for Support Vector Machines [J]. Pattern recognition, 2006, 39(7): 1333-1345.
- [11] Narasimha M J, Peterson A M. On The Computation of The Discrete Cosine Transform [J]. Communications, IEEE Transactions on, 1978, 26(6): 934-936.
- [12] Schölkopf B, Smola A, Müller K R. Kernel Principal Component Analysis [M]. Berlin Springer, 1997: 583-588.
- [13] Shensa M J. The Discrete Wavelet Transform: Wedding The a Trous and Mallat Algorithms [J]. Signal Processing, IEEE Transactions on, 1992, 40(10): 2464-2482.
- [14] Yan F, Mikolajczyk K, Barnard M, et al. l p Norm Multiple Kernel Fisher Discriminant Analysis for Object and Image Categorisation [C]. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010: 3626-3632.

数据理解与建模

一些博客资源对本章的组织给予了很多启示，如 <http://www.cnblogs.com/> 中的 @jerrylead、@tornadomeet、@liqizhou、@hrlin 等，以及 <http://blog.csdn.net/> 中的 @jinshengtao、@jedial_lu 等，在此对各位博主表示感谢。本书的润色得到了南京大学计算机科学与技术系及智能信息处理研究组的张雷博士及蒋澜、李红、王茜、陈厚兵、陆恒杨、李明、张文哲等几位同学的协助，在此表示深深的谢意。

8.1 引言

作为五氏之一的神农氏出生在烈山的一个石洞里，天生异象：身体透明、牛头人身……

天赋异禀且勤劳勇敢的神农氏长大后被推举为部落首领，称为炎帝。炎帝之名据说是因为其领导的部落处于炎热的南部；神农氏之称，是因为他发明了五谷农业……

神农氏以耨鞭鞭草木，始尝百草……（摘自西汉《史记》）

神农以耨鞭鞭百草，尽知其平、毒、寒、温之性，臭味所主……（摘自晋代干宝的《搜神记》）

神农尝百药之时，“……皆口尝而身试之，一日之间而遇七十毒……其所得三百六十物……后世承传为书，谓之《神农本草（经）》。”（摘自宋代郑樵的《通志》）

事实上，炎帝和神农氏是否是同一个人暂时还没有明确的定论，比如：《史记·五帝本纪》隐喻炎帝与神农氏并非一人；《史记·封禅书》分列炎帝和神农氏为两人；而《世本·帝系篇》则首次将炎帝和神农氏合在一起称为“炎帝神农氏”；汉高诱注《淮南子·时则训》也提及赤帝即炎帝，少典之子，号为神农，南方火德之帝；东汉郑玄注《礼记》和三国韦昭注《国语》均提及烈山氏为炎帝；《水经注》卷三十二也将烈山氏和神农氏相并……

本章无意就炎帝和神农氏是否是同一个人的历史公案作任何的论证和说明，本章关注的是在漫长的人类历史上，人们在社会实践中不断总结出一系列被相信或经过检验而被认可的规律和知识，用于指导日常生活和生产，以《神农本草经》为例，全书共记载了 365 种药物，

其中植物药 252 种、动物药 67 种、矿物药 46 种；除了从气、味、毒性、药效等各个角度将每种药物进行分类之外（《神农本草经》记载：“药有酸、咸、甘、苦、辛五味，又有寒、热、温、凉四气，及有毒、无毒……采治时月生熟，土地所出，真伪陈新，并各有法”），该书还最早提出了“七情”学说，即所谓单行、相须、相使、相畏、相杀、相恶、相反，基本涵盖了中药配伍的相关内容，其理论意义对当代中药学研究的影响依旧巨大。基于神农氏的开创之举，后来的历代医家不断地完善和丰富，经过几千年的积淀，形成了如今名扬中外的中华医学国粹——中医。

从数据分析的角度来看神农氏的伟大工作，其在中医药领域的贡献（包括但不限于）如表 8-1 所示。

表 8-1 数据分析师视角下的“神农尝百草”

序号	神农氏之于中医	数据分析师的视角
1	通过亲自口尝百草，从品类繁多的植物中分拣出草药（除了草药外，传说神农氏也分拣出可食用的农作物，此处不表）	通过相关性分析（“植物-疗效”）去除原始数据中与目标（疗效）无关的噪声
2	通过亲自口尝百草，利用自身的体验，对于每一种草药，从气、味、毒的角度提取了各种草药的特征	（利用感官）标注不同实体对象（草药）不同维度的特征
3	将不同的草药用于不同的病患，根据其疗效，得到每一种草药的功效	通过先验数据（应该是多次实践）计算出“（每个）药材-疗效”的先验概率
4	通过将不同的草药进行组合以用于不同的病患，根据其疗效，得到不同草药组合在一起所形成的药方	1) 通过先验数据（应该是多次实践）计算出“（多个）药材-疗效”的联合概率 2) 特征选择：从 365 种药材中选择与具体疗效相关的有效组合
5	通过将不同的草药进行不同的配伍组合，根据其疗效，得出药与药之间的配伍原则和配伍禁忌	1) （应该是通过多次实践）计算出属性（每一味药材）之间的相关性 2) （应该是通过多次实践）在不同的上下文场景中对待药材进行了分类（君药、臣药、佐药、使药）
...

更为形式化的数学描述如下（以中药方剂为例）：

- 1) 中药方剂描述的是“药-证”的关系模型，表示如下：
- $$y_j = f(x_i), \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, M$$
- 其中， x_i 表示每一种中草药， i 表示药物的序号； y_j 表示每一种证候， j 表示证候的序号（中医讲究辨证施治，事实上也可以用功效来表示，此处以证候作为示例）。
- 2) 辨证表示的是“望闻问切-证候”的关系模型，表示如下：
- $$z_j = g_j(t_i), \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, M$$
- 其中， t_i 表示通过望闻问切所获得的反映病人体征的属性数据， i 表示属性的序号； z_j 表示每一种证候， j 表示证候的序号。
- 我们的先人通过若干年若干次的实践，建立起了形如 $y_j = f(x_i)$ 或 $z_j = g_j(t_i)$ 的模型，不

过古人建立的模型是经验性质的，并没有一种数学的描述，而是一系列的示例集或规则集（有用，但未必完备、也未必完全一致）。中医的“辨证施治”类似于人工智能里的 CBR 推理（基于示例的推理），推理步骤可以简单描述为：

1) 辨证：通过对当前病人进行望闻问切所获得的属性数据与历史数据相比，判断当前病人的属性数据对应的是哪一个类别（分类问题，类别对应于证候的类别）。

2) 施治：罗列所有类似当前证候的历史数据（古方、经方等），以历史上当前证候的给药方案作为当前病人的给药依据。

如何将以示例和经验驱动的“辨证施治”形式化为数学模型是中医客观化研究的重要一环，从数据分析师的角度来看，如何从既有的示例数据中建立“望闻问切数据 - 证”“药 - 效”“量 - 效”关系是一个典型的数据建模问题。数据建模的一个基本假设是：数据中存在一种模式（我们尚未发现，但一定存在）。

这种模式可能是产生这些数据的“上帝”在产生这些数据时所依赖的规律和准则，或者说，有那么一个潜在的“上帝”隐秘地“操纵”这些对象实体按照某一个（组）规律或准则产生数据。而所谓的从数据中发掘规律、知识和洞见的本质目标就是发现由所谓的“上帝”显式或隐式地彰显的旨意（规律和准则）。为了下文表述的方便，我们将这些唯有“上帝”才能知晓并使用的规律和准则记为 f ，则所谓的从大数据中发掘知识和洞见，指的就是根据可观测到的既有数据建立一个映射函数 f' ，目标是希望 f' 和“上帝”的旨意 f 一致，并通过某些评价指标来评估 f' 。

如何从既有数据中利用某种手段找到与 f 尽可能一致的 f' 就是典型的数据挖掘（或者知识发现）问题。可以简单地将数据挖掘看成是数据（库）和机器学习的合集，前者关注数据如何存取，而后者关注于如何利用已有的数据进行建模，即从既有数据中拟合出那个反映“上帝”旨意的函数。

本章将对用于数据建模的机器学习方法进行概要的介绍，并以十大经典机器学习方法为主例，阐述各个算法的数学原理并给出应用示例和提示，本章下面的结构安排如下：8.2 节扼要介绍机器学习与数据建模的基本定义和分类；8.3 节简单介绍非监督学习及典型的非监督学习算法；8.4 节从回归和分类两个角度介绍监督学习及典型的监督算法；8.5 节对本章进行小结。

8.2 机器学习

在心理学理论中，学习是指（人或动物）依靠经验的获得而使行为持久变化的过程。在机器学习的场景下，不同的学者有不同的理解和定义。比如，西蒙（Simon）认为：如果一个系统能够通过执行某种过程而改进它的性能，这就是学习。明斯基（M. Minsky）认为：学习是在人们头脑中（心理内部）进行有用的变化。汤姆·米切尔（Tom M. Mitchell）认为：对于某类任务 T 和性能度 P ，如果一个计算机程序在 T 上以 P 衡量的性能随着经验 E 而自我完善，

那么，我们称这个计算机程序在从经验 E 中学习。

根据不同的分类准则，机器学习又可以分为不同的类别，具体参见表 8-2。

表 8-2 不同分类准则意义下的机器学习

序号	分类准则	机器学习类别	功能描述
1	数据对象	符号学习	机器学习的数据对象是非数值型数据，分为记忆学习、示例学习、决策树学习、演绎学习和类比学习
		数值学习	机器学习的数据对象是数值型数据
2	学习方法	归纳学习	基于归纳的一种学习方法，操作步骤包括泛化和特化，目标是从给定的样本中归纳出一个一般的概念描述
		演绎学习	基于演绎推理的一种学习，学习系统由给定的知识进行演绎的保真推理，并存储有用的结论
		类别学习	基于类比推理的学习方法，寻找和利用事物间可类比的关系，从已有的知识中推导出未知的知识
		分析学习	基于数学分析的学习方法
3	学习目标	概念学习	学习的目标和结果为概念，是为了获得概念的学习
		规则学习	学习的目标和结果为规则，是为了获得规则的学习
		函数学习	学习的目标和结果为函数，是为了获得函数的学习
		类别学习	学习的目标和结果为对象类，是为了获得类别的学习
		贝叶斯学习	学习的目标和结果是贝叶斯网络，是为了获得贝叶斯网络的一种学习
4	学习方式	监督学习	利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程，也称为监督训练或有导师学习
		非监督学习	用于处理未被分类标记的样本集，输入数据中无导师信号，采用聚类方法，学习结果为类别
		强化学习	以环境反馈（奖/惩罚信号）作为输入，以统计和动态规划技术为指导的一种学习
5	学习策略	模拟人脑	模拟人脑的宏观心理级或微观生理级学习过程，分为符号学习和连接学习两类，前者基于认知心理学，后者基于脑和神经科学原理
		数学方式	以样本数据为依据，以概率统计理论为基础，以数值运算为方法的一类机器学习
6	数据形式	结构化学习	以结构化数据为输入，基于数值计算或符号推演的一种学习
		非结构化学习	以非结构化数据为输入，基于数值计算或符号推演的一种学习

事实上，具体到每一种机器学习方法，根据上述不同的分类准则，可能会归属到一个或多个类别中。

2006 年 12 月，ICDM（the IEEE International Conference on Data Mining）邀请了机器学习与数据挖掘领域的顶尖专家从候选的 18 个机器学习算法中评选出了十大经典机器学习算法。虽然这个评选已经过去 10 年，但这十大算法仍然被公认为是机器学习领域中的经典，俨然成

为进入机器学习与数据挖掘领域必备的技能。这十大算法（按排名顺序）分别是：

- 1) C4.5 算法。
- 2) K-Means 算法。
- 3) SVM 算法。
- 4) Apriori 算法。
- 5) EM 算法。
- 6) PageRank 算法。
- 7) AdaBoost 算法。
- 8) KNN 算法。
- 9) NB 算法。
- 10) CART 算法。

需要注意的是，这 18 个候选算法也是由顶尖专家提名的，能进入提名的，事实上都是经典，除了后来选出的十大算法之外，另外的 8 个候选算法分别是：

- 1) FP-Tree 算法。
- 2) HITS 算法。
- 3) BIRCH 算法。
- 4) GSP 算法。
- 5) PreFixSpan 算法。
- 6) CBA 算法。
- 7) Finding Reduct 算法。
- 8) gSpan 算法。

下章将围绕监督学习及非监督学习着重介绍上述相关算法和方法。

8.3 非监督学习

在非监督学习（Unsupervised Learning）中，数据并不会被特别标识，学习模型是为了推断出数据的一些内在结构。非监督学习一般有两种思路：

1) 第一种思路是在指导 Agent 时不为其指定明确的分类，而是在成功时采用某种形式的激励制度。需要注意的是，这类训练通常会被置于决策问题的框架里，因为它的目标不是产生一个分类系统，而是做出最大回报的决定，这类学习往往被称为强化学习。

2) 第二种思路称为聚合（Clustering），这类学习类型的目标不是让效用函数最大化，而是找到训练数据中的近似点，本节将重点介绍此类非监督学习思路。

第二种思路的非监督学习常见的应用场景包括关联规则的学习及聚类等。常见算法包括 Apriori、K-Means、EM 等，由于 4.3.3 节中已经介绍过 Apriori，因此下面将重点介绍后两者。

8.3.1 K-Means

K-Means 算法是典型的基于距离的聚类算法, K-Means 认为:

- 1) 两个对象的距离越近, 其相似度就越大。
- 2) 相似度接近的若干对象可组成一个聚集 (也可称为“簇”)。
- 3) K-Means 的目标是从给定数据集中找到紧凑且独立的簇。

K-Means 中的 K 指的就是在数据集中找出的聚集 (“簇”) 的个数, 在 K-Means 算法中, 此 K 的大小需要事先设定, K-Means 的算法流程如下:

输入: 数据集, K

输出: K 个聚集

Step1: 从 N 个数据对象中任意选择 K 个对象作为初始聚类中心, 记为 $u_j (j = 1, 2, \dots, K)$ 。

Step2: 根据每个聚类对象的均值 (中心对象), 计算每个对象与这些中心对象的距离, 并根据最小距离重新对相应的对象进行划分, 即

$$C_i = \operatorname{argmin}_j \|x_i - u_j\|^2$$

Step3: 重新计算每个 (有变化) 聚类的均值 (中心对象), 即

$$u_j = \frac{\sum_{i=1}^m 1\{c_i = j\} \cdot x_i}{\sum_{i=1}^m 1\{c_i = j\}}$$

Step4: 循环 Step2 到 Step3 直到每个聚类不再发生变化 (收敛) 为止。

K-Means 聚类算法的优点集中体现在 (包括但不限于):

- 1) 算法快速、简单。
- 2) 对大数据集有较高的计算效率并且可伸缩。
- 3) 时间复杂度近于线性, 适合挖掘大规模数据集。

K-Means 聚类算法的时间复杂度是 $O(N \cdot K \cdot T)$, 其中 N 代表数据集中对象的数量; T 代表算法迭代的次数; K 代表簇的数目; 一般而言: $K \ll N$ 且 $T \ll N$ 。

K-Means 的缺陷集中体现在 (包括但不限于):

- 1) 在 K-Means 算法中, K 是事先设定的, 而 K 值的选定是非常难以估计的。很多时候, 事先并不知道给定的数据集应该被分成多少个类别才最合适。
- 2) 在 K-Means 算法中, 初始聚类中心的选择对聚类结果有较大的影响, 一旦初始值选择得不好, 就可能无法得到有效的聚类结果。
- 3) K-Means 算法需要不断地进行样本分类调整, 不断地计算调整后的新的聚类中心, 因此当数据量非常大时, 算法的时间开销是非常大的。

8.3.2 EM

在统计计算中,最大期望(EM)算法是在依赖无法直接观测的隐藏变量(Latent Variable)的概率模型(Probabilistic Model)中寻找参数最大似然估计或最大后验估计的算法,最大期望经常用于机器学习和计算机视觉的数据聚类(Data Clustering)领域。

将一瓶酒均分到两个杯子的过程是描述EM算法的形象例子,这个例子的目标是将一瓶酒(重量不知)均分到两个没有刻度的(透明)杯子中(每个杯子的均分目标是1/2瓶)。在没有其他辅助设备(比如天平或量杯)的情况下,一个简单的做法是先把酒随意地倒入两个杯子中,然后观测是否一样多,把比较多的一杯倒出一些到另外一个较少的杯子中,然后再观察,再将比较多的一杯倒出一些到另外一个较少的杯子,这个过程一直迭代地进行,直到在视觉上看不出两个杯子的酒有什么分量上的不同为止。这个例子有如下几个要点:

- 1) 我们无法知道每个杯子是否倒入了1/2瓶,但是只要知道其中一个杯子倒入了多少,我们就可以知道另外一个杯子倒入了多少。
- 2) 迭代倒酒的目标是两个杯子均分一瓶酒,迭代过程的核心过程有两步:第一步是随意地在两个杯子中倒入酒;第二步是调整每个杯子中的酒(将多的倒入少的杯子)。
- 3) 迭代收敛(终止)的条件是:在视觉上看不出两个杯子的酒有什么分量上的不同,或者说,已经无法在肉眼上区分两杯分量的不同,或者在迭代动作上无法转移(从多的杯子到少的杯子)更细微的酒量。上述这3个特征是EM算法的3个要素。

EM算法就是这样,假设要估计 A 和 B 两个参数,在开始状态下二者都是未知的,并且知道了 A 的信息就可以得到 B 的信息,反之亦然(如上例中知道其中一个杯子的量就能知道另外一个杯子的量)。可以考虑首先赋予 A 某种初始值,以此得到 B 的估计值,然后从 B 的当前值出发,重新估计 A 的取值,这个过程一直持续到收敛为止(收敛条件将在下文详述)。

一般地,用 Y 表示随机变量的观测数据, Z 表示隐随机变量的数据。 Y 和 Z 组合在一起成为完全数据,观测数据 Y 又称为不完全数据。假设给定观测数据,其概率分布是 $P(Y|\theta)$,其中 θ 是需要估计的概率模型参数,那么不完全数据 Y 的似然函数是 $P(Y|\theta)$,对数似然函数是 $L(\theta) = \log P(Y|\theta)$,假设 Y 和 Z 的联合概率分布是 $P(Y, Z|\theta)$,则其完全数据的对数似然函数是 $L(\theta) = \log P(Y, Z|\theta)$,求解 θ 就是极大似然估计。

一般地,“似然”常常被用作作为“概率”的同义词,但是在统计学中,二者具有截然不同的用法:“概率”描述了已知参数时的随机变量的输出结果;“似然”则用来描述已知随机变量的输出结果时,未知参数的可能取值。例如,对于“一枚正反对称的硬币上抛10次”这种事件,我们可以问硬币落地时10次都是正面向上的“概率”是多少;而对于“一枚正反对称的硬币上抛10次,落地都是正面向上”的这种事件,我们则可以问这枚硬币正反面对称的“似然”程度是多少。

极大似然估计是参数估计的方法之一,是概率论在统计学中的应用,指的是已知某个随机样本满足某种概率分布,但是其中的具体参数不清楚,参数估计就是通过若干次实验,观测其结果,利用其结果计算出参数的大概值。极大似然原理的直观想法是:在一次随机实验中,结果 A 出现,则一般认为实验条件对 A 有利(即 A 出现的概率很大),一般而言,若事件 A 发生的概率与参数 θ 相关, A 发生的概率记为 $P(A, \theta)$,则 θ 的估计应该使上述概率达到最大,这样的 θ 被称为极大似然估计。求极大似然估计的一般步骤是:①写出似然函数;②对似然函数取对数;③求导数;④解似然方程。

一般而言,极大似然估计很难有解析解,只有通过迭代的方法求解。EM 算法就是用于求解极大似然估计的一种迭代算法。EM 算法通过迭代求 $L(\theta) = \log P(Y, Z | \theta)$ 的极大似然估计,每次迭代包含两步:第一步求期望,称为 E 步骤;第二步求极大化,称为 M 步骤。EM 算法描述如下:

输入:观测变量数据 Y 、隐变量数据 Z 、联合分布 $P(Y, Z | \theta)$ 、条件分布 $P(Z | Y, \theta)$

输出:模型参数 θ

Step1: 设置参数的初值 $\theta^{(0)}$ 。

Step2: 记 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计值,在第 $i+1$ 次迭代的 E 步,计算:

$$Q(\theta | \theta^{(i)}) = E_z(\log P(Y, Z | \theta) | Y, \theta^{(i)}) = \sum_z \log P(Y, Z | \theta) P(Z | Y, \theta^{(i)})$$

其中, $P(Z | Y, \theta^{(i)})$ 是在给定观测数据 Y 和当前的参数估计 $\theta^{(i)}$ 下隐变量数据 Z 的条件概率分布。

Step3: 求使 $Q(\theta | \theta^{(i)})$ 极大化的 θ , 确定第 $i+1$ 次迭代的参数的估计值 $\theta^{(i+1)}$:

$$\theta^{(i+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(i)})$$

Step4: 重复 Step2 (称为 E 步) 和 Step3 (称为 M 步), 直到收敛。

从上述的 EM 算法描述中,可以看出, $Q(\theta | \theta^{(i)})$ 是核心,称为 Q 函数。所谓 Q 函数,指的是完全数据的对数似然函数 $L(\theta) = \log P(Y, Z | \theta)$ 关于在给定观测数据 Y 和当前参数 $\theta^{(i)}$ 下对未观测数据 Z 的条件概率分布 $P(Z | Y, \theta^{(i)})$ 的期望,即

$$Q(\theta | \theta^{(i)}) = E_z(\log P(Y, Z | \theta) | Y, \theta^{(i)})$$

关于 EM 算法的几点说明如下:

- 1) Step1 步的参数初值 $\theta^{(0)}$ 可以设置为任意值,但 EM 算法对初值是敏感的。
- 2) Step2 步 (E 步) 中的 $Q(\theta | \theta^{(i)})$ 中的第一个变量 θ 表示要极大化的参数,第二个参数 $\theta^{(i)}$ 表示当前参数的估计值,每次迭代实际都是在求 Q 函数及其极大值。
- 3) Step3 步 (M 步) 求 $Q(\theta | \theta^{(i)})$ 的极大化,得到 $\theta^{(i+1)}$,完成一次迭代。
- 4) Step4 步给出迭代的条件,一般是对较小的正数 ε_1 及 ε_2 ,若满足下述条件即停止迭代:

$$|\theta^{(i+1)} - \theta^{(i)}| < \varepsilon_1 \text{ 或 } |Q(\theta | \theta^{(i+1)}) - Q(\theta | \theta^{(i)})| < \varepsilon_2$$

为了更好地理解上述流程,本文以一个简单的实例(选自《统计学习方法》,略有修改)介绍具体的实现过程。假设有3枚硬币,分别记作A、B和C。这些硬币正面出现的概率分别为 π 、 p 和 q 。投币实验如图8-1所示,先投A,如果A是正面,即 $A=1$,那么选择投B;如果 $A=0$,那么选择投C。最后,如果B或C是正面,那么 $Y=1$;如果是反面,那么 $Y=0$ 。

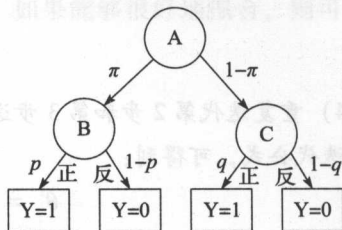


图8-1 三硬币实验

独立重复 N 次实验($N=10$),观测结果如表8-3所示。

表8-3 三硬币实验结果

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
1	1	0	1	0	0	1	0	1	1

假设只能观测到投掷硬币(B或C)的结果输出,而不能观测到投掷硬币的过程。问如何估计三枚硬币正面出现的概率,即求 π 、 p 和 q 的值。

上述实验中的三硬币模型可以写作:

$$P(y|\theta) = \sum_z P(y, z|\theta) = \sum_z P(z|\theta)P(y|z, \theta) = \pi p^y (1-p)^{1-y} + (1-\pi)q^y (1-q)^{1-y}$$

其中,随机变量 y 是观测变量(可观测),表示每一次的观测结果(1或0);随机变量 z 是隐变量(不可观测),表示观测到的投硬币A的结果; $\theta=(\pi, p, q)$ 是参数模型。

将观测数据表示为 $Y=(y_1, y_2, \dots, y_N)^T$,未观测变量数据表示为 $Z=(z_1, z_2, \dots, z_N)$,则观测数据的似然函数为:

$$P(Y|\theta) = \sum_Z P(Z|\theta)P(Y|Z, \theta) = \prod_{j=1}^N (\pi p^{y_j} (1-p)^{1-y_j} + (1-\pi)q^{y_j} (1-q)^{1-y_j})$$

利用EM算法求解思路如下:

1) 设置初始值,记作: $\theta^0=(\pi^0, p^0, q^0)$,然后通过E步骤和M步骤迭代计算参数的估计,直到收敛为止。令第 i 次迭代参数的估计值为 $\theta^i=(\pi^i, p^i, q^i)$,EM算法的第 $i+1$ 次迭代如下第2步和第3步所述,第2步是E步骤,第3步是M步骤。

2) 计算在模型参数 $\theta^i=(\pi^i, p^i, q^i)$ 下,观测数据 y_j 来自投硬币B的概率:

$$u_j^{i+1} = \frac{\pi^i (p^i)^{y_j} (1-p^i)^{1-y_j}}{\pi^i (p^i)^{y_j} (1-p^i)^{1-y_j} + (1-\pi^i) (q^i)^{y_j} (1-q^i)^{1-y_j}}$$

3) 计算模型参数的新估计值,计算公式如下:

$$\pi^{(i+1)} = \frac{1}{N} \sum_{j=1}^N u_j^{i+1}$$

$$p^{(i+1)} = \frac{\sum_{j=1}^N u_j^{i+1} y_j}{\sum_{j=1}^N u_j^{i+1}}$$

$$q^{(i+1)} = \frac{\sum_{j=1}^N (1 - u_j^{i+1}) y_j}{\sum_{j=1}^N (1 - u_j^{i+1})}$$

4) 重复迭代第2步和第3步进行计算。假设模型参数的初值取为 $\theta^0 = (0.5, 0.5, 0.5)$, 利用迭代公式, 可得到:

$$\begin{aligned}\theta^1 &= (\pi^1, p^1, q^1) = (0.5, 0.6, 0.6) \\ \theta^2 &= (\pi^2, p^2, q^2) = (0.5, 0.6, 0.6)\end{aligned}$$

于是得到模型参数的极大似然估计 $\theta = (0.5, 0.6, 0.6)$ 。

如果取初值 $\theta^0 = (0.4, 0.6, 0.7)$, 那么得到的模型参数的极大似然估计值是 $\theta = (0.4064, 0.5368, 0.6432)$, 由此可见 EM 算法与初值的选择有关, 选择不同的初值可能得到不同的参数估计值。

8.4 监督学习

监督学习 (Supervised Learning) 是指利用一组已知明确标识或结果的样本调整分类器的参数, 使其达到所要求性能的过程, 也称为有教 (导) 师学习。所谓“监督”或“有教 (导) 师”指的是监督学习必须依赖一个已经明确标记的训练数据 (训练集) 作为监督学习的输入 (学习素材)。训练集由若干个训练实例组成, 每个实例都是一个属性集合 (通常为向量, 代表对象的特征) 和一个明确的标识 (可以是离散的, 也可以是连续的) 组成。监督学习的过程就是建立预测模型的学习过程, 将预测结果与训练集的实际结果进行比较, 不断地调整预测模型, 直到模型的预测结果达到一个预期的准确率。

根据训练集中的标识是连续的还是离散的, 可以将监督学习分为两类: 回归和分类。前者对应于训练集的标识是连续的情况, 而后者适用于训练集的标识是离散的场景, 离散的标识往往称为类标 (label)。

8.4.1 回归

回归是研究一个随机变量 Y 或一组随机变量 $Y(y_1, y_2, \dots, y_n)$ 对一个属性变量 X 或一组属性变量 $X(x_1, x_2, \dots, x_n)$ 的相依关系的统计分析方法, 通常称 X 或 $X(x_1, x_2, \dots, x_n)$ 为自变量, 称 Y 或 $Y(y_1, y_2, \dots, y_n)$ 为因变量。如果自变量个数大于 1, 则称为多元回归, 如果因变量个数大于 1, 则称为多重回归。回归分析的结果是得到一类数学模型 (函数)。当因变量和自变量的关系是线性时, 则称为线性模型 (这是最简单的一类数学模型), 更为特殊的情况是自变量和因变量均只有一个, 且它们大体上呈线性关系, 这类回归分析被称为一元线性回归。当数学模型的函数形式是未知参数的线性函数时, 称为线性回归模型; 当函数形式是未知参数的非线性函数时, 称为非线性回归模型。

回归分析的一般过程是通过因变量和自变量建立回归模型,并根据训练集求解模型的各个参数,然后评价回归模型是否能很好地拟合测试集实例,如果能够很好地拟合,则可以根据自变量进行因变量的预测,回归分析的主要步骤是:

- 1) 寻找 h 函数 (即 hypothesis)。
- 2) 构造 $J(W)$ 函数 (又称损失函数)。
- 3) 调整参数 W 使得 $J(W)$ 函数最小。

1. 线性回归

线性回归模型假设自变量 (也称输入特征) 和因变量 (也称目标值) 满足线性关系。为了便于叙述,取自变量为 $X(x_1, x_2, \dots, x_n)$, 因变量为 Y , 训练参数为 $W(w_1, w_2, \dots, w_n)$ 。

- 1) 目标数学模型函数定义为:

$$Y(W, X) = w_1x_1 + w_2x_2 + \dots + w_nx_n$$

- 2) 基于最小二乘定义损失函数为:

$$J(W) = \frac{1}{2} \sum_{i=1}^m (Y_i - W^T X_i)^2$$

其中 X_i 和 Y_i 分别表示训练集中第 i 个样本的自变量和因变量, m 表示训练集的个数, 前面乘上系数 $(1/2)$ 是为了求导的时候, 使常数系数消失。

- 3) 调整参数 W 使得 $J(W)$ 最小, 即

$$\frac{dJ(W)}{dW} = 0$$

具体的方法有梯度下降法、最小二乘法等, 下面先以梯度下降法介绍求解思路。对 W 取一个随机初始值, 然后不断地迭代改变 W 的值使得 J 减小, 直到最终收敛 (取得一个 W 值使得 $J(W)$ 最小)。 W 的迭代更新规则如下:

$$W_{j+1} = W_j - \varepsilon \frac{\partial}{\partial W_j} J(W)$$

其中, ε 称为学习率 (Learning Rate), j 表示 W 的迭代次数, 将 $J(W)$ 代入上式得到的更新公式是:

$$W_{j+1} = W_j - \varepsilon \frac{\partial}{\partial W_j} J(W) = W_j - \frac{\varepsilon}{2} \frac{\partial}{\partial W_j} \sum_{i=1}^m (Y_i - W_j^T X_i)^2 = W_j + \varepsilon \sum_{i=1}^m (Y_i - W_j^T X_i) X_i$$

此更新规则称为 LMS (Least Mean Squares, 最小均方, 通常不用汉译形式) 更新策略, 也称为 Widrow-Hoff learning rule, 从上述更新公式可以看到, W 的每一次迭代都要考察训练集的所有样本, 这种更新策略称为批量梯度下降 (batch gradient descent)。还有一种更新策略是随机梯度下降 (stochastic gradient descent), 其基本思路是每处理一个训练样本就更新一次 W 。相比较而言, 由于批量梯度下降在每一步都考虑全部数据集, 因而复杂度比较高; 随机梯度下降的收敛会比较快。在实际情况中两种梯度下降得到的最优解 $J(W)$ 一般都会接近真实的最小值, 所以对于较大的数据集, 一般采用效率较高的随机梯度下降法。

为了便于理解上述的计算流程, 以一个具体的示例加以说明, 示例设置如下:

示例训练集: $X(x_1, x_2, x_3) = (2, 5, 3)$, $Y = 850$

示例模型函数: $Y = w_1x_1 + w_2x_2 + w_3x_3$

学习速率: $\varepsilon = \frac{1}{35}$ (人为设置)

整个训练过程中各个参数的变化如表 8-4 所示, 为了便于阅读, 将每次迭代 W 的变化罗列在表中, 即表中的 $\Delta\omega_1$ 、 $\Delta\omega_2$ 、 $\Delta\omega_3$ 。

表 8-4 简单迭代过程示意

次数	ω_1	ω_2	ω_3	Error	$\Delta\omega_1$	$\Delta\omega_2$	$\Delta\omega_3$
1	50.00	50.00	50.00	350.00	20.00	50.00	30.00
2	70.00	100.00	80.00	-30.00	-1.71	-4.29	-2.57
3	68.29	95.71	77.43	2.57	0.15	0.37	0.22
4	68.43	96.08	77.65	-0.22	-0.01	-0.03	-0.02
5	68.42	96.05	77.63	0.02	—	—	—

为了表示方便, 表 8-4 中的数值均保留 2 位小数, 并且仅显示了 5 步迭代的计算过程 (假定 0.02 是可以接受的误差), 从表 8-4 可以看出, 经过 5 步迭代后可得到的回归模型函数是:

$$Y = 68.42x_1 + 96.05x_2 + 77.63x_3$$

事实上, 对于形如 $X(x_1, x_2, \dots, x_n) = (2, 5, 3)$, $Y = 850$ 的样本, 其模型或许是 $Y = 150x_1 + 50x_2 + 100x_3$, 或许是 $Y = 100x_1 + 100x_2 + 50x_3$, 这意味着如下两点:

- 1) 从回归的角度而言, 结果可能并不唯一。
- 2) 回归结果未必是数据样本本来的模型。

对于后者, 如果有更多的学习样本, 或许会有利于让结果更加逼近训练集背后的模型, 这或许也是大数据时代, 为什么要更热衷于“大”的数据的原因, 大概是因为, 唯有以更“大”的数据作为支撑, 才有可能发掘数据背后的那个知识或洞见 (模型)。

刚才提及的更新策略是梯度下降法, 需要多次迭代, 相对来说比较费时而且也不太直观。除了梯度下降法之外, 还有最小二乘法更新策略。最小二乘法的计算思路是基于矩阵论, 将权值的计算从梯度下降法的迭代改为矩阵计算, 经过推导可以知道:

$$W = (X^T X)^{-1} X^T Y$$

限于篇幅原因, 此处不做具体的推导。无论是梯度下降法还是最小二乘法, 其在拟合的过程中都是基于这样的假设: $X(x_1, x_2, \dots, x_n)$ 中“每一个属性的重要性 (权重) 都是一样的”, 而这在实际场景中未必适用 (往往会产生过拟合或欠拟合的现象), 针对这种情况就产生了加权的线性回归思路, 其本质是对各个元素进行规范化处理, 对不同的输入特征赋予不同的非负值权重, 权重越大, 对于代价函数的影响就越大。

特别值得一提的是: 上述提到的线性回归模型 $Y(W, X) = w_1x_1 + w_2x_2 + \dots + w_nx_n$, 所谓的线性是对参数 W 而言的, 并非一定是输入 $X(x_1, x_2, \dots, x_n)$ 的线性函数, 比如可以通过一

系列的基函数 $\varphi_i(\cdot)$ 对输入进行非线性变换, 即

$$Y(W, X) = \sum_{i=1}^n w_i \varphi_i(X)$$

其中, $\varphi_i(\cdot)$ 是基函数, 可选择的基函数有多项式、高斯函数、Sigmoid 函数等, 简单介绍如下:

(1) 多项式

多项式函数是由常数与自变量经过有限次乘法与加法运算得到的, 定义如下:

$$\varphi(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$$

其中, $a_i (i=0, 1, \cdots, n)$ 是常数, 当 $n=1$ 时, 多项式函数为一次函数 $\varphi(x) = a_1 x + a_0$ 。

(2) 高斯函数

高斯函数的定义如下:

$$\varphi(x) = a \cdot \exp\left(-\frac{(x-b)^2}{c^2}\right)$$

其中, a 、 b 及 c 均是实常数, 且 $a > 0$ 。

(3) Sigmoid 函数

Sigmoid 函数是一个在生物学中常见的 S 函数, 定义如下:

$$\varphi(x) = \frac{1}{1 + e^{-x}}$$

2. Logistic 回归

Logistic 回归一般用于分类问题, 而其本质是线性回归模型, 只是在回归的连续值结果上加了一层函数映射, 将特征线性求和, 然后使用 $g(z)$ 进行映射, 将连续值映射到一个区间内, 然后在该区间内取定一个阈值作为分类边界, 示意图如图 8-2 所示。根据映射函数 $g(z)$ 的不同选择, 其

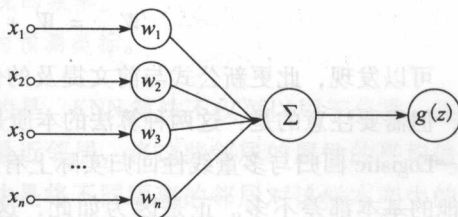


图 8-2 Logistic 回归模型示意

分类性能也不同, 比如, 如果映射函数是 Sigmoid 函数时, 则其分类结果为“0”和“1”两类, 而如果映射函数是双曲正弦 \sinh 函数时, 其分类结果则为“1”和“-1”两类。

以 Sigmoid 二值化 (Sigmoid 函数的特征是: 当自变量趋于 $-\infty$, 因变量趋近于 0, 而当自变量趋近于 ∞ , 因变量趋近于 1) 为例, 为了便于后文的叙述, 将 $Y(W, X)$ 写作 $h_w(X)$, 图 8-2 所示的 Logistic 回归模型如下:

$$h_w(X) = Y(W, X) = g(W^T X) = \frac{1}{1 + e^{-W^T X}}$$

由于输出目标是离散值, 与上述的线性回归思路应该有所不同, 令

$$P(Y = 1 | X, W) = h_w(X)$$

$$P(Y = 0 | X, W) = 1 - h_w(X)$$

上述两个公式可以统一写作:

$$P(Y|X, W) = (h_w(X))^Y (1 - h_w(X))^{1-Y}$$

则似然估计是:

$$L(W) = P(Y|X, W) = \prod_{j=1}^n P(Y_j|X_j, W) = \prod_{j=1}^n (h_w(X_j))^{Y_j} (1 - h_w(X_j))^{1-Y_j}$$

其中 j 表示训练集中的样本序号, 为了便于求解, 对 $L(W)$ 进行取对数操作, 即

$$l(W) = \log(L(W)) = \sum_{i=1}^n (Y_i \log(h_w(X_i)) + (1 - Y_i) \log(1 - h_w(X_i)))$$

接下来是 W 的似然估计最大化, 可以考虑上述的梯度下降法, 于是得到:

$$\frac{\partial l(W)}{\partial W_j} = \sum_{i=1}^n (Y_i \frac{1}{g(W_j^T X_i)} - (1 - Y_i) \frac{1}{1 - g(W_j^T X_i)}) \frac{\partial}{\partial W_j} g(W_j^T X_i)$$

由于

$$\frac{\partial}{\partial W_j} g(W_j^T X_i) = g(W_j^T X_i) (1 - g(W_j^T X_i)) \frac{\partial}{\partial W_j} W_j^T X_i$$

所以, 上式可以进一步简化为:

$$\frac{\partial l(W)}{\partial W_j} = \sum_{i=1}^n (Y_i (1 - g(W_j^T X_i)) - (1 - Y_i) g(W_j^T X_i)) X_i = \sum_{i=1}^n (Y_i - g(W_j^T X_i)) X_i$$

因此, 更新公式是:

$$W_{j+1} = W_j + \varepsilon \sum_{i=0}^m (Y_i - h_{W_j}(X_i)) X_i$$

可以发现, 此更新公式与前文提及的使用梯度下降法求解线性回归而采用的迭代公式相同, 但需要注意的是: 这两种算法的本质是不一样的。

Logistic 回归与多重线性回归实际上有很多相同之处, 最大的区别就是它们的因变量不同, 其他的基本都差不多。正是因为如此, 这两种回归可以归于同一个家族, 即广义线性模型 (generalized linear model)。Logistic 回归的因变量可以是二分类的, 也可以是多分类的, 但是二分类的更为常用, 也更加容易解释。所以实际中最为常用的就是二分类的 Logistic 回归。如果因变量是多分类的, 则扩展为 Softmax 回归。Softmax 回归模型是 logistic 模型在多分类问题上的推广, 在 Softmax 回归中, 类标签 Y 可以取 $k(k > 2)$ 个不同的值, 其推导思路与 Logistic 回归相同, 本文不再赘述。

8.4.2 分类

分类问题是机器学习研究中的一个重要问题, 与回归问题类似, 分类过程也是从训练集中建立因变量和自变量的映射过程。与回归问题不同的是, 在分类问题中, 因变量的取值是离散的, 根据因变量的取值范围 (个数) 可将分类问题分为二分类问题 (比如“好人”或“坏人”)、三分类问题 (比如“支持”、“中立”或“反对”) 及多分类问题。在分类问题中, 因变量称为类标 (label), 而自变量称为属性 (或特征)。

根据分类采用的策略和思路的不同,分类算法包括(但不限于):基于示例的分类方法(代表算法是KNN)、基于概率模型的分类方法(代表算法是朴素贝叶斯、最大期望算法EM)、基于线性模型的分类方法(代表算法是SVM)、基于决策模型的分类方法(代表算法包括:C4.5、AdaBoost、随机森林)等,下面简单介绍上述各种典型的分类算法的问题背景和算法思路。

1. KNN

K最近邻(K-Nearest Neighbor, KNN)分类算法是一个在理论方面比较成熟的方法,也是最简单的机器学习算法之一。该方法的出发点是:如果一个样本在特征空间中的 k 个最相似(即特征空间中的最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别,并具有这个类别上样本的特性。KNN算法是从训练集中找到和新数据最接近的 k 条记录,然后根据他们的主要类别来决定新数据的类别。该算法涉及3个主要因素:训练集、距离或相似的度量、 k 的大小。算法的执行步骤如下:

输入:训练集(包括 n 个已经标注的记录 (X, Y))、 k 、测试用例 $X(x_1, x_2, \dots, x_n)$

输出:测试用例的类标

Step1: 遍历训练集中的每个记录,计算每个记录的属性特征 $X_i(i=1, 2, \dots, n)$ 与测试用例 $X(x_1, x_2, \dots, x_n)$ 的距离,记为 $D_i(i=1, 2, \dots, n)$ 。

Step2: 从 $D_i(i=1, 2, \dots, n)$ 中选择最小的 k 个记录(样本)。

Step3: 统计这 k 个记录(样本)对应的类别出现的频率。

Step4: 返回出现频率最高的类别作为测试用例的预测类标。

KNN的思想很好理解,也很容易实现。更重要的是,KNN算法不仅可以用于分类,还可以用于回归,具体思路是通过找出一个样本的 k 个最近邻居,将这些邻居的属性的平均值赋给该样本,就可以得到该样本的属性。更有用的方法是将不同距离的邻居对该样本产生的影响给予不同的权值(如权值与距离成反比),使得回归更加普适。KNN算法的不足之处在于:

1) 每次分类都需要和训练集中的所有记录进行一次距离或相似度的计算,如果训练集很大,则计算负担很重。

2) 从上述记录流程中可以看出,如果 k 个近邻的类别属性各异,则会给分类带来麻烦(需要其他策略支持)。

2. 朴素贝叶斯

朴素贝叶斯分类是利用统计学中的贝叶斯定理来预测类成员的概率,即给定一个样本,计算该样本属于一个特定的类的概率,朴素贝叶斯分类基于的一个假设是:每个属性之间都是相互独立的,并且每个属性对分类问题产生的影响都是一样的。

贝叶斯定理由英国数学家贝叶斯(Thomas Bayes)发展而来,用于描述两个条件概率之间的关系,比如 $P(A|B)$ 和 $P(B|A)$,其中 $P(A|B)$ 表示在事件 B 已经发生的前提下,事件 A 发生的概率,称为事件 B 发生下事件 A 的条件概率,其基本公式是:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

按照乘法法则:

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

由上式推导可以得到:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

一座别墅在过去的 20 年里一共发生过 2 次被盗, 别墅的主人有一条狗, 狗平均每周晚上叫 3 次, 在盗贼入侵时狗叫的概率被估计为 0.9, 问题是: 在狗叫的时候发生入侵的概率是多少?

用贝叶斯的理论求解此问题, 假设 A 事件为“狗在晚上叫”, B 为“盗贼入侵”, 则:

$$1) P(A) = \frac{3}{7} \quad (\text{计算根据: 狗平均每周晚上叫 3 次});$$

$$2) P(B) = \frac{2}{20 \times 365} \quad (\text{计算根据: 过去 20 年发生过 2 次被盗});$$

$$3) P(A|B) = 0.9 \quad (\text{计算根据: B 事件发生时 A 事件发生的概率是 0.9}).$$

基于上述数据, 可以很容易地计算出 A 事件发生时 B 事件发生的概率 $P(B|A)$ 是:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} = \frac{0.9 \times \frac{2}{20 \times 365}}{\frac{3}{7}} = 0.00058$$

朴素贝叶斯分类的出发点是: 对于给出的待分类项, 求解在出现此项的条件下各个类别出现的概率, 哪个最大, 就认为此待分类项属于哪个类别。为了便于描述, 将事件 A 表示为特征属性 $X(x_1, x_2, \dots, x_n)$, 将事件 B 表示为类标属性 $Y(y_1, y_2, \dots, y_m)$, 则朴素贝叶斯分类问题可以描述为: 对于一个给定的测试样本的特征属性 $X(x_1, x_2, \dots, x_n)$, 求其属于各个类标 $y_i (i=1, 2, \dots, m)$ 的概率 $P(y_i|X)$ 中的最大值, 基于前面的定义可以知道:

$$P(y_i|X) = \frac{P(X|y_i)P(y_i)}{P(X)}$$

其中 X 表示特征属性 (x_1, x_2, \dots, x_n) , 由于朴素贝叶斯是基于属性独立性的假设 (前文已提及), 故

$$P(X|y_i) = P(x_1|y_i)P(x_2|y_i)\cdots P(x_n|y_i) = \prod_{j=1}^n P(x_j|y_i)$$

$$P(y_i|X) = \frac{\prod_{j=1}^n P(x_j|y_i)P(y_i)}{P(X)}$$

又由于 $P(X)$ 是一个常数, 因此只要比较分子的大小即可。朴素贝叶斯分类器的算法流程如下:

输入：训练集，测试用例 $X(x_1, x_2, \dots, x_n)$

输出：测试用例 $X(x_1, x_2, \dots, x_n)$ 的类标

Step1: 遍历训练集，统计各个类别下各个特征属性的条件概率估计，即

$$P(x_i | y_j) (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$$

Step2: 遍历训练集，根据上述公式，计算 $P(y_1 | X)$, $P(y_2 | X)$, \dots , $P(y_m | X)$ 。

Step3: 如果 $P(y_k | X) = \max\{P(y_1 | X), P(y_2 | X), \dots, P(y_m | X)\}$ ，则测试用例的类标是 y_k 。

为了更好地理解上述计算流程，以一个具体的实例来说明。已知一个训练集如表 8-5 所示，特征属性有两个，分别是 color 和 weight，其中，color 的取值范围是 $\{0, 1, 2, 3\}$ ；weight 的取值范围是 $\{0, 1, 2, 3, 4\}$ 。类标属性有 1 个 (sweet)，取值范围是 $\{\text{yes}, \text{no}\}$ 。

表 8-5 训练集示意

color	weight	sweet
3	4	yes
2	3	yes
0	3	no
3	2	no
1	4	no

测试用例是 (color = 3; weight = 4)，求其类标？

遍历训练集，可以得到：

$$P(y_1) = P(y = \text{yes}) = \frac{2}{5}; \quad P(y_2) = P(y = \text{no}) = \frac{3}{5}$$

遍历训练集，可以得到：

$$P(x_1 = 0 | y_1) = 0; \quad P(x_1 = 1 | y_1) = 0; \quad P(x_1 = 2 | y_1) = \frac{1}{2}; \quad P(x_1 = 3 | y_1) = \frac{1}{2};$$

$$P(x_2 = 0 | y_1) = 0; \quad P(x_2 = 1 | y_1) = 0; \quad P(x_2 = 2 | y_1) = 0; \quad P(x_2 = 3 | y_1) = \frac{1}{2};$$

$$P(x_2 = 4 | y_1) = \frac{1}{2};$$

$$P(x_1 = 0 | y_2) = \frac{1}{3}; \quad P(x_1 = 1 | y_2) = \frac{1}{3}; \quad P(x_1 = 2 | y_2) = 0; \quad P(x_1 = 3 | y_2) = \frac{1}{3};$$

$$P(x_2 = 0 | y_2) = 0; \quad P(x_2 = 1 | y_2) = 0; \quad P(x_2 = 2 | y_2) = \frac{1}{3}; \quad P(x_2 = 3 | y_2) = \frac{1}{3};$$

$$P(x_2 = 4 | y_2) = \frac{1}{3}$$

因为测试用例是 (color = 3; weight = 4)，所以

$$P(y_1 | (x_1 = 3; x_2 = 4)) \propto P(x_1 = 3 | y_1) P(x_2 = 4 | y_1) P(y_1) = \frac{1}{2} \times \frac{1}{2} \times \frac{2}{5} = \frac{1}{10};$$

$$P(y_2 | (x_1 = 3; x_2 = 4)) \propto P(x_1 = 3 | y_2)P(x_2 = 4 | y_2)P(y_2) = \frac{1}{3} \times \frac{1}{3} \times \frac{3}{5} = \frac{1}{15}$$

由于 $P(y_1 | (x_1 = 3; x_2 = 4)) > P(y_2 | (x_1 = 3; x_2 = 4))$, 故测试用例的类标是 y_1 , 即 yes。

通过上述的计算实例可以发现, 事实上, 把 $P(x_i | y_j)$ 的所有可能均事先计算出来是没有必要的, 只需要根据测试用例的具体样本进行选择性的计算即可。

理论上, 朴素贝叶斯分类模型与其他分类方法相比具有最小的误差率, 但其独立性假设在实际应用中往往是不成立的, 这给朴素贝叶斯分类模型的正确分类带来了一定的影响。针对这个缺点, 有一些改进的算法, 此处不作罗列。

3. SVM

支持向量机 (Support Vector Machines, SVM) 是一种面向二分类问题的监督学习方法, SVM 的基本思路是将向量映射到一个更高维的空间里, 然后在此空间里构建特征空间中间隔最大的线性分类器 (从而将数据分离成两部分)。因此, SVM 的学习策略就是间隔最大化, 可形式化为一个求解凸二次规划的问题。针对训练数据是否线性可分, 可以将 SVM 分为以下 3 类:

- 1) 线性可分支持向量机: 当训练数据线性可分时, 通过硬间隔最大化, 学习一个线性的分类器, 该线性分类器称为线性可分支持向量机, 又称为硬间隔支持向量机。
- 2) 线性支持向量机: 当训练数据接近线性可分时, 通过软间隔最大化, 学习一个线性的分类器, 该线性分类器称为线性支持向量机, 又称为软间隔支持向量机。
- 3) 非线性支持向量机: 当训练数据线性不可分时, 通过使用核技巧及软间隔最大化, 学习一个非线性支持向量机。

当输入空间为欧氏空间或离散集合, 特征空间为希尔伯特空间时, 核函数表示将输入从输入空间映射到特征空间得到的特征向量之间的内积。通过使用核函数可以隐式地在高维的特征空间中学习线性支持向量机, 这样的方法称为核技巧。

(1) 线性可分支持向量机

假设给定一个特征空间上的训练数据集:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

其中, $x_i \in R^n$, $y_i \in \{+1, -1\}$, $i = 1, 2, \dots, N$ 。

如上所述的训练集中, (x_i, y_i) 称为样本点, x_i 称为第 i 个特征向量 (也称为实例), y_i 是 x_i 的类标记。当 $y_i = +1$ 时, 称 x_i 为正例; 当 $y_i = -1$ 时, 称 x_i 为负例。

二分类的目标是在特征空间中找到一个可分离超平面 (分类器), 将实例分到不同的类 (正类和负类), 这个超平面对应的方程是 $w^T x + b = 0$ 。当训练数据集线性可分时, 存在无穷个这样的超平面将两类数据正确分开, 如图 8-3 所示。

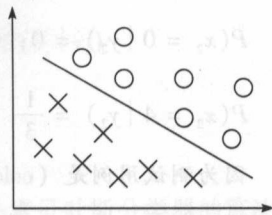


图 8-3 二分类问题

对于类似图 8-3 所示的二分类问题,事实上可以有很多直线将两类准确无误地分离开来,线性可分支持向量机对应着将两类数据正确划分开来并且使两个间隔超平面之间距离最大的直线,显然,这样的直线是唯一的,我们将此超平面记为 $w^T x + b = 0$, 对应的分类决策函数是 $f(x) = \text{sign}(w^T x + b)$ 。

对于给定的训练数据集 T 和超平面 (w, b) , 定义超平面 (w, b) 关于样本点 (x_i, y_i) 的函数间隔为:

$$\hat{\gamma}_i = y_i(w^T x_i + b)$$

超平面 (w, b) 关于训练数据 T 的函数间隔为超平面 (w, b) 关于 T 中所有样本点 (x_i, y_i) 的函数间隔的最小值, 即

$$\hat{\gamma} = \min_{i=1,2,\dots,N} \hat{\gamma}_i$$

在超平面 $w^T x + b = 0$ 确定的情况下, 对于任意一点 (x_0, y_0) , $|w^T x_0 + b|$ 能够相对地表示点 x_0 距离超平面的远近, 而 $w^T x_0 + b$ 的符号与类标记 y_0 的符号是否一致可以表示分类是否正确。所以可用 $y(w^T x + b)$ 表示分类的正确性及确信度, 这就是函数间隔的定义动机。

函数间隔可以表示分类预测的正确性及确信度, 但是在选择分离超平面时, 只有函数间隔还不够, 比如只要成比例地改变 w 和 b , 超平面并没有改变, 但函数间隔却成比例地缩放了的, 因此还需要其他的约束条件。

对于给定的训练数据集 T 和超平面 (w, b) , 定义超平面 (w, b) 关于样本点 (x_i, y_i) 的几何间隔为:

$$\gamma_i = y_i \left(\frac{w^T}{\|w\|} x_i + \frac{b}{\|w\|} \right) = \frac{\hat{\gamma}_i}{\|w\|}$$

超平面 (w, b) 关于训练数据集 T 的几何间隔为超平面 (w, b) 关于 T 中所有样本点 (x_i, y_i) 的几何间隔的最小值, 即

$$\gamma = \min_{i=1,2,\dots,N} \gamma_i = \frac{\hat{\gamma}}{\|w\|}$$

为训练数据集找到几何间隔最大化的超平面意味着以充分大的确信度对训练数据进行分类。最大间隔分离超平面可以表示为 $\max_{w,b} \gamma$ 的约束最优化问题, 约束条件是:

$$y_i \left(\frac{w^T}{\|w\|} x_i + \frac{b}{\|w\|} \right) \geq \gamma, \quad i = 1, 2, \dots, N$$

鉴于几何间隔和函数间隔的关系, 上述优化问题可以改写为 $\gamma = \max_{w,b} \frac{\hat{\gamma}}{\|w\|}$ 的约束最优化问题, 约束条件为:

$$y_i(w^T x_i + b) \geq \hat{\gamma}, \quad i = 1, 2, \dots, N$$

函数间隔 $\hat{\gamma}$ 的取值并不影响最优化问题的解, 因此, 为了简化上述问题, 可以取 $\hat{\gamma} = 1$ 。另外, 最大化 $\frac{1}{\|w\|}$ 和最小化 $\frac{\|w\|^2}{2}$ 是等价的, 因此上述的优化问题可以等价于 $\min_{w,b} \frac{\|w\|^2}{2}$ 的约束优化

问题, 约束条件是:

$$y_i(w^T x_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N$$

在线性可分的情况下, 训练数据集的样本点中与分离超平面距离最近的样本点的实例称为支持向量, 支持向量便是满足 $y_i(w^T x_i + b) - 1 = 0$ 的点, 如图 8-4 所示。

如图 8-4 所示, 中间的实线便是寻找到的最优超平面, 其到两条虚线之间的距离相等, 这个距离便是几何间隔 $\hat{\gamma}$, 两条虚线之间的距离等于 $2\hat{\gamma}$, 而虚线上的点则是支持向量。由于这些支持向量刚好在边界上, 所以它们满足 $y_i(w^T x_i + b) - 1 = 0$ 。

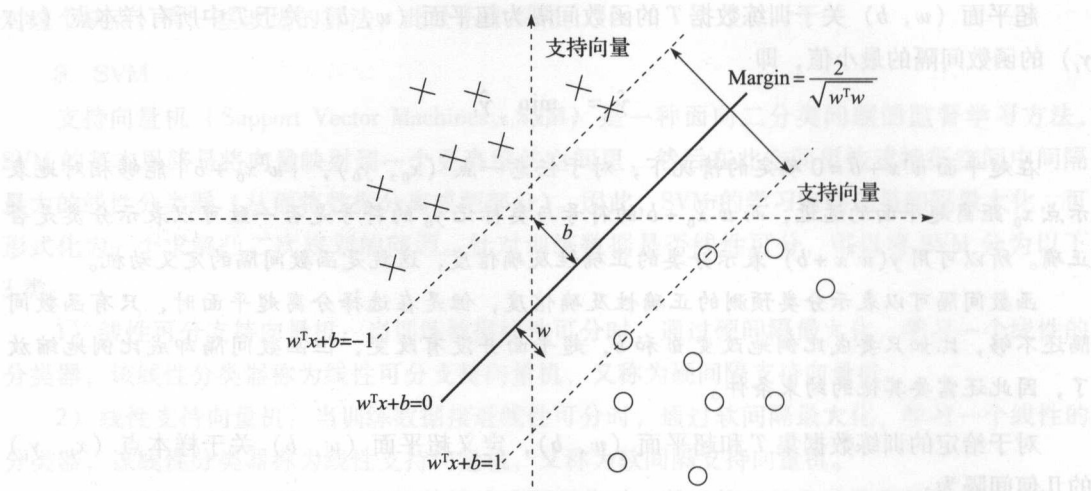


图 8-4 支持向量示意

在决定分离超平面时只有支持向量起作用, 而其他实例点并不起作用。如果移动支持向量将改变所求的解。但是如果在间隔边界以外移动其他实例点, 甚至去掉这些点, 解是不会改变的。由于支持向量在确定分离超平面中起着决定性的作用, 所以将这种分类模型称为“向量机”。支持向量的个数一般很少, 所以支持向量机是由很少的“重要的”训练样本确定的。

以一个具体的实例为例加以描述, 如图 8-5 所示的训练数据集, 其正例是 $x_1 = (3, 3)^T$ 、 $x_2 = (4, 3)^T$, 负例是 $x_3 = (1, 1)^T$, 试求最大间隔分离超平面。

依据训练数据集构造约束最优化问题为 $\min_{w,b} \frac{\|w\|^2}{2} =$
 $\min_{w,b} \frac{w_1^2 + w_2^2}{2}$, 约束条件是:

$$\begin{cases} 3w_1 + 3w_2 + b \geq 1, & (1) \\ 4w_1 + 3w_2 + b \geq 1, & (2) \\ -w_1 - w_2 - b \geq 1, & (3) \end{cases}$$

由约束条件 (1) 和 (3) 可得 $w_1 + w_2 \geq 1$, 所以当 $w_1 =$

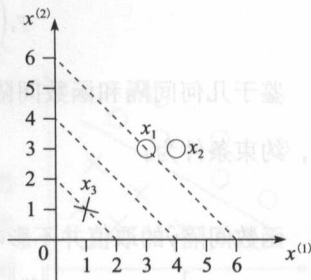


图 8-5 SVM 实例

$w_2 = \frac{1}{2}$ 时, 可取得最优值, 此时 $b = -2$, 最大分离超平面是 $\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$, 其中, $x_1 = (3, 3)^T$ 和 $x_3 = (1, 1)^T$ 是支持向量。

为了求解线性可分支持向量机的最优化问题, 应用拉格朗日对偶性, 通过求解对偶问题得到原始问题的最优解, 这就是线性可分支持向量机的对偶算法。这样做的优点在于:

- 1) 对偶问题往往更容易求解。
- 2) 自然引入核函数, 进而推广到非线性分类问题。

具体步骤如下所述:

首先构建拉格朗日函数, 然后对每一个不等式约束引进拉格朗日乘子 $a_i \geq 0 (i = 1, 2, \dots, N)$, 定义拉格朗日函数如下:

$$L(w, b, a) = \frac{\|w\|^2}{2} - \sum_{i=1}^N a_i y_i (w^T x_i + b) + \sum_{i=1}^N a_i$$

其中, $a = (a_1, a_2, \dots, a_N)^T$ 为拉格朗日乘子向量, 根据拉格朗日对偶性, 原始问题的对偶问题是极大极小问题, 即 $\max_a \min_{w, b} L(w, b, a)$, 求解步骤如下:

- 1) 求解 $\min_{w, b} L(w, b, a)$ 。将 $L(w, b, a)$ 对 w 和 b 分别求偏导数并令其等于 0, 得到:

$$\begin{cases} \frac{\partial L(w, b, a)}{\partial w} = w - \sum_{i=1}^N a_i y_i x_i = 0 \\ \frac{\partial L(w, b, a)}{\partial b} = \sum_{i=1}^N a_i y_i = 0 \end{cases} \Rightarrow \begin{cases} w = \sum_{i=1}^N a_i y_i x_i \\ \sum_{i=1}^N a_i y_i = 0 \end{cases}$$

代入原拉格朗日函数, 可得:

$$L(w, b, a) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) - \sum_{i=1}^N a_i y_i \left(\left(\sum_{j=1}^N a_j y_j x_j \right)^T x_i + b \right) + \sum_{i=1}^N a_i$$

化简得

$$L(w, b, a) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) + \sum_{i=1}^N a_i$$

$$\text{即 } \min_{w, b} L(w, b, a) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) + \sum_{i=1}^N a_i$$

- 2) 求对偶问题, 即 $\max_a \left(-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) + \sum_{i=1}^N a_i \right)$, 约束条件是:

$$\sum_{i=1}^N a_i y_i = 0, \quad a_i \geq 0, \quad i = 1, 2, \dots, N$$

上述优化问题等价于 (求极大转换成求极小) $\min_a \left(\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) - \sum_{i=1}^N a_i \right)$, 约束条件是:

$$\sum_{i=1}^N a_i y_i = 0, \quad a_i \geq 0, \quad i = 1, 2, \dots, N$$

设 $a^* = (a_1^*, a_2^*, \dots, a_i^*)^T$ 是上述对偶最优问题的解, 则存在下标 j , 使得 $a_j^* \geq 0$, 并按下式求得原始最优优化问题的解:

$$\begin{cases} w = \sum_{i=1}^N a_i^* y_i x_i \\ b = y_j - \sum_{i=1}^N a_i^* y_i (x_i^T x_j) \end{cases}$$

仍以上面的实例为例, 假设一训练数据集, 其正例是 $x_1 = (3, 3)^T$ 、 $x_2 = (4, 3)^T$, 负例是 $x_3 = (1, 1)^T$, 试求最大间隔分离超平面。

数据集的对偶问题是 $\min_a \left(\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) - \sum_{i=1}^N a_i \right)$, 代入得到的对偶问题是:

$$\min_a \left(\frac{1}{2} (18a_1^2 + 25a_2^2 + 2a_3^2 + 42a_1a_2 - 12a_1a_3 - 14a_2a_3) - a_1 - a_2 - a_3 \right)$$

其约束条件是: $a_1 + a_2 - a_3 = 0$, $a_i \geq 0$, $i = 1, 2, 3$ 。

将 $a_1 + a_2 - a_3 = 0$, $a_i \geq 0$, $i = 1, 2, 3$ 代入目标函数可得:

$$s(a_1, a_2) = 4a_1^2 + \frac{13}{2}a_2^2 + 10a_1a_2 - 2a_1 - 2a_2$$

对 a_1 和 a_2 求偏导并令其为 0, 得

$$\begin{cases} \frac{\partial s(a_1, a_2)}{\partial a_1} = 8a_1 + 10a_2 - 2 = 0 \\ \frac{\partial s(a_1, a_2)}{\partial a_2} = 13a_2 + 10a_1 - 2 = 0 \end{cases} \Rightarrow a_1 = \frac{3}{2}, a_2 = -1$$

由于 $a_2 = -1$, 不满足 $a_2 \geq 0$ 的约束条件, 表明最小值应该在边界上:

1) 若 $a_1 = 0$, 那么代入 $\frac{\partial s(a_1, a_2)}{\partial a_2} = 13a_2 + 10a_1 - 2 = 0$, 得到 $a_2 = \frac{2}{13}$, 将 a_1 和 a_2 代入 $s(a_1, a_2)$, 得到 $s\left(0, \frac{2}{13}\right) = -\frac{2}{13}$ 。

2) 若 $a_2 = 0$, 那么代入 $\frac{\partial s(a_1, a_2)}{\partial a_1} = 8a_1 + 10a_2 - 2 = 0$, 得到 $a_1 = \frac{1}{4}$, 将 a_1 和 a_2 代入 $s(a_1, a_2)$, 得到 $s\left(\frac{1}{4}, 0\right) = -\frac{1}{4}$ 。

由于 $s\left(\frac{1}{4}, 0\right) < s\left(0, \frac{2}{13}\right)$, 即 $a_1 = \frac{1}{4}$, $a_2 = 0$ 时达到最小, 此时 $a_3 = a_1 + a_2 = \frac{1}{4}$ 。

因此 $a_1^* = a_3^* = \frac{1}{4}$ 对应的 x_1 和 x_3 是支持向量。代入得到:

$$\begin{cases} w = \sum_{i=1}^N a_i^* y_i x_i = \frac{1}{4} \cdot (1) \cdot (3, 3)^T + \frac{1}{4} \cdot (-1) \cdot (1, 1)^T = \left(\frac{1}{2}, \frac{1}{2}\right)^T \\ b = y_j - \sum_{i=1}^N a_i^* y_i (x_i^T x_j) = 1 - \frac{1}{4} \cdot (1) \cdot ((3, 3)^T (3, 3)) - \frac{1}{4} \cdot (-1) \cdot ((1, 1)^T (3, 3)) \\ = -2 \end{cases}$$

所以最大分离超平面是 $\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$ 。

(2) 线性支持向量机

假设给定一个特征空间上的训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

其中, $x_i \in R^n$, $y_i \in \{+1, -1\}$, $i = 1, 2, \dots, N$ 。

如上所述的训练集中, (x_i, y_i) 称为样本点, x_i 称为第 i 个特征向量 (也称为实例), y_i 是 x_i 的类标记。假设训练数据集不是线性可分的, 通常情况下, 训练数据中有一些特异点, 将这些特异点除去后, 剩下大部分样本点组成的集合就是线性可分的。

线性不可分意味着某些样本点 (x_i, y_i) 不能满足函数间隔大于等于 1 的约束条件, 为了解决这个问题, 可以对每个样本 (x_i, y_i) 引进一个松弛变量 $\xi_i \geq 0$, 使得函数间隔加上松弛变量大于等于 1, 这样, 约束条件就变为:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N$$

同时, 对每个松弛变量 ξ_i , 支付一个代价 ξ_i , 目标函数由原来的 $\frac{\|w\|^2}{2}$ 变为

$$\frac{\|w\|^2}{2} + C \sum_{i=1}^N \xi_i, \quad C > 0$$

其中, C 为惩罚参数, 一般根据应用场景自行设定。线性不可分的线性支持向量机的学习问题就变成了凸二次规划问题 (原始问题), 即寻优 $\min_{w, b, \xi} \left(\frac{\|w\|^2}{2} + C \sum_{i=1}^N \xi_i \right)$, 约束条件是:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N$$

设问题的解是 w 和 b , 则分离超平面是 $w^T x + b = 0$, 其对应的分类决策函数是 $f(x) = \text{sign}(w^T x + b)$, 这样的模型称为训练样本线性不可分时的线性支持向量机, 简称为线性支持向量机, 显然线性支持向量机包含线性可分支持向量机。由于现实中的训练数据集往往是线性不可分的, 线性支持向量机具有更广的适用性。

原始最优化问题的拉格朗日函数是:

$$L(w, b, \xi, a, u) = \frac{\|w\|^2}{2} - C \sum_{i=1}^N \xi_i - \sum_{i=1}^N a_i (y_i (w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^N u_i \xi_i$$

其中, $a_i \geq 0$, $u_i \geq 0$, 对偶问题是拉格朗日极大极小值问题。首先求 $L(w, b, \xi, a, u)$ 对 w , b 和 ξ 的极小值, 可得到:

$$\begin{cases} \frac{\partial L(w, b, \xi, a, u)}{\partial w} = w - \sum_{i=1}^N a_i y_i x_i = 0 \\ \frac{\partial L(w, b, \xi, a, u)}{\partial b} = - \sum_{i=1}^N a_i y_i = 0 \\ \frac{\partial L(w, b, \xi, a, u)}{\partial \xi} = C - a_i - u_i = 0 \end{cases} \Rightarrow \begin{cases} w = \sum_{i=1}^N a_i y_i x_i \\ \sum_{i=1}^N a_i y_i = 0 \\ C - a_i - u_i = 0 \end{cases}$$

将结果代入 $L(w, b, \xi, a, u)$ 可得:

$$\min_{w, b, \xi} (L(w, b, \xi, a, u)) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) + \sum_{i=1}^N a_i$$

求其对偶问题 $\max_a \left(-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) + \sum_{i=1}^N a_i \right)$, 约束条件为:

$$\begin{cases} \sum_{i=1}^N a_i y_i = 0 \\ C - a_i - u_i = 0 \\ a_i \geq 0 \\ u_i \geq 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^N a_i y_i = 0 \\ 0 \leq a_i \leq C \end{cases}$$

再将对目标函数求极大转化为求极小, 可得到等价的对偶问题:

$$\min_a \left(\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i^T x_j) - \sum_{i=1}^N a_i \right)$$

其约束条件是:

$$\begin{cases} \sum_{i=1}^N a_i y_i = 0 \\ 0 \leq a_i \leq C \end{cases} \quad i = 1, 2, \dots, N$$

考虑图 8-6a 中的两种类别, A 点似乎偏离 Class 2 有点远, 如果我们直接忽略它, 原来的分隔超平面还是挺好的, 但是由于这个特异点的出现, 导致分隔超平面不得不被挤歪了, 同时两类之间的间隔也相应变小了。当然, 更严重的情况是, 如果出现图 8-6b 中的这种特异点, 我们将无法构造出能将数据线性分开的超平面来。

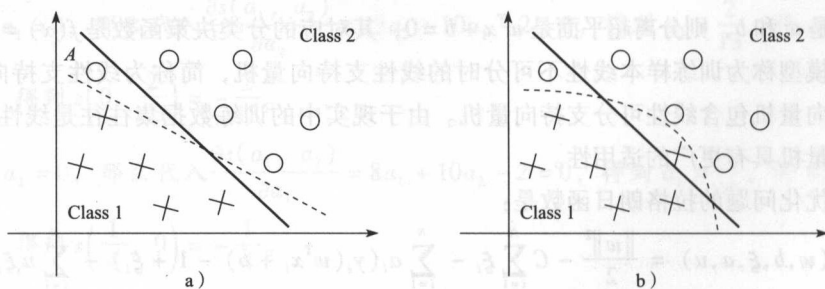


图 8-6 二类问题中的超平面

为了处理这种情况, 我们允许数据点在一定程度上偏离超平面。也就是允许一些点跑到 H_1 和 H_2 之间, 即它们到分类面的间隔会小于 1, 如图 8-7 所示。

(3) 非线性支持向量机

非线性分类的问题是指利用非线性模型才能很好地进行分类的问题。如图 8-8 所示, 该数据集不是线性可分的。

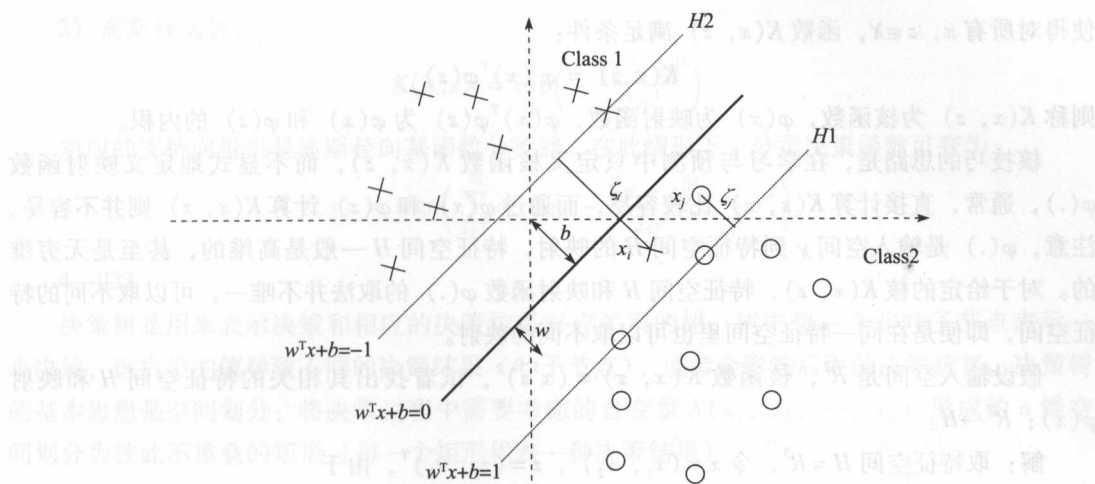


图 8-7 间隔小于 1 的超平面

假设给定一个特征空间上的训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

其中, $x_i \in R^n, y_i \in \{+1, -1\}, i = 1, 2, \dots, N$ 。

如上所述的训练集中, (x_i, y_i) 被称为样本点, x_i 被称为第 i 个特征向量 (也称为实例), y_i 是 x_i 的类标记。如果能用 R^n 中的一个超平面将正负例正确地分离开, 则称这个问题是非线性可分问题。

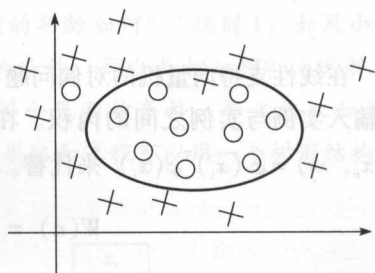


图 8-8 非线性分类问题

非线性问题往往不好求解, 所以希望能用解线性分类问题的方法来解决这个问题。所采取的方法是进行一个非线性变换, 将非线性问题变换为线性问题, 通过解决变换后的线性问题的方法解决原来的非线性问题。

设原空间为 $X \subset R^2, X = (x_1, x_2)^T \in X$;

设新空间为 $Z \subset R^2, z = (z_1, z_2)^T \in Z$;

定义从原空间到新空间的变换 (映射) 是:

$$z = \varphi(x) = (x_1^2, x_2^2)^T$$

经过 (非线性) 变换 $z = \varphi(x)$, 原空间 $X \subset R^2$ 变换成新空间 $Z \subset R^2$, 原空间的点也相应地变换为新空间的点。

用线性分类方法求解非线性分类问题的过程分为两步: 首先使用一个变换将原空间的数据映射到新空间; 然后在新空间里利用线性分类学习方法从训练数据中学习分类模型。

设 X 是输入空间, H 为特征空间, 如果存在一个从 X 到 H 的映射:

$$\varphi(x): X \rightarrow H$$

使得对所有 $x, z \in \mathcal{X}$, 函数 $K(x, z)$ 满足条件:

$$K(x, z) = \varphi(x)^T \varphi(z)$$

则称 $K(x, z)$ 为核函数, $\varphi(x)$ 为映射函数, $\varphi(x)^T \varphi(z)$ 为 $\varphi(x)$ 和 $\varphi(z)$ 的内积。

核技巧的思路是, 在学习与预测中只定义核函数 $K(x, z)$, 而不显式地定义映射函数 $\varphi(\cdot)$, 通常, 直接计算 $K(x, z)$ 比较容易, 而通过 $\varphi(x)$ 和 $\varphi(z)$ 计算 $K(x, z)$ 则并不容易。注意, $\varphi(\cdot)$ 是输入空间 \mathcal{X} 到特征空间 H 的映射, 特征空间 H 一般是高维的, 甚至是无穷维的。对于给定的核 $K(x, z)$, 特征空间 H 和映射函数 $\varphi(\cdot)$ 的取法并不唯一, 可以取不同的特征空间, 即便是在同一特征空间里也可以取不同的映射。

假设输入空间是 R^2 , 核函数 $K(x, z) = (x^T z)^2$, 试着找出其相关的特征空间 H 和映射 $\varphi(x): R^2 \rightarrow H$ 。

解: 取特征空间 $H = R^3$, 令 $x = (x_1, x_2)^T$, $z = (z_1, z_2)^T$, 由于

$$(x^T z)^2 = (x_1 z_1 + x_2 z_2)^2 = (x_1 z_1)^2 + 2x_1 z_1 x_2 z_2 + (x_2 z_2)^2$$

所以可以取映射:

$$\varphi(x) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)^T$$

在线性支持向量机的对偶问题中, 无论是目标函数还是决策函数 (分离超平面) 都只涉及输入实例与实例之间的内积。在求对偶问题的目标函数中的内积 $x_i^T x_j$ 时, 可以用核函数 $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ 来代替。此时, 对偶函数的目标函数可变为:

$$W(a) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N a_i$$

同样, 分类决策函数中的内积也可以用核函数来代替, 而分类决策函数式变为:

$$f(x) = \text{sign} \left(\sum_{i=1}^{N_s} a_i^* y_i \varphi(x_i)^T \varphi(x) + b^* \right) = \text{sign} \left(\sum_{i=1}^{N_s} a_i^* y_i K(x_i, x) + b^* \right)$$

这等价于经过映射函数 $\varphi(\cdot)$ 将原来的输入空间变换到一个新的特征空间, 将输入空间中的内积 $x_i^T x_j$ 变换到特征空间中的内积 $\varphi(x_i)^T \varphi(x_j)$, 在新的特征空间里从训练样本中学习线性支持向量机。当映射函数是非线性函数时, 学习到的含有核函数的支持向量机是非线性分类模型。

也就是说, 在核函数 $K(x_i, x_j)$ 给定的条件下, 可以利用解线性分类问题的方法求解非线性分类问题的支持向量机。学习是在特征空间里隐式地进行的, 不需要显式地定义特征空间和映射函数。这样的技巧称为核技巧, 它是巧妙地利用线性分类学习方法与核函数解决非线性问题的技术。常见的核函数有:

1) 多项式核函数:

$$K(x, z) = (\gamma x^T z + c)^p$$

对应的支持向量机是一个 p 次多项式分类器。在此情形下, 分类决策函数可变为:

$$f(x) = \text{sign} \left(\sum_{i=1}^{N_s} a_i^* y_i (\gamma x_i^T x + c)^p + b^* \right)$$

2) 高斯核函数:

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

对应的支持向量机是高斯径向基函数分类器。在此情形下, 分类决策函数可变为:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N_s} a_i^* \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) + b^*\right)$$

4. ID3

决策树是用来表示决策和相应的决策结果对应关系的树, 树中每一个非叶子节点表示一个决策, 该决策的值导致不同的决策结果 (叶子节点), 或者会影响后面的决策选择。决策树的基本思想是空间划分: 将决策过程中需要考虑的自变量 $X(x_1, x_2, \dots, x_n)$ 形成的 n 维空间划分为彼此不重叠的矩形 (每一个矩形代表一种决策结果)。

决策树的思想并不复杂, 事实上, 我们每天进行的很多决策和判断都是基于此思路进行的。以男女双方相亲的场景为例, 从女方的角度而言, 其可能的决策思路 (本示例仅用于解释决策树技术, 不代表价值观取向) 和流程是: ①这个男孩的年龄如何? (规则1: 如果小于等于30岁就约会, 否则不予约会); ②如果年龄符合见面的条件, 那么长相如何? (规则2: 如果长相比较帅气就约会, 否则不予约会); ③如果长相符合见面的条件, 那么收入如何? (规则3: 如果收入高就约会, 否则不予约会)。上面的决策思路和流程可以用一个树形结构来表示, 如图8-9a所示。

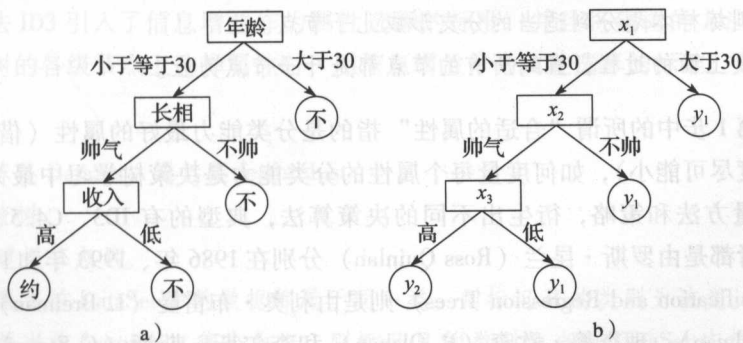


图8-9 决策树示例图

参见图8-9a, 可以看到: 该决策考虑的因素 (自变量) 有三个, 分别是“年龄”“长相”和“收入”, 用 $X(x_1, x_2, x_3)$ 表示, 其中 x_1 代表“年龄”、 x_2 代表“长相”、 x_3 代表“收入”; 该决策的决策结果 (因变量) 有两个, 分别是“约会”和“不约”, 并且全部是叶子节点, 用 $Y(y_1, y_2)$ 表示, 其中 y_1 代表“不约”、 y_2 代表“约会”。事实上, 上述的决策流程可以理解为针对 $X(x_1, x_2, x_3)$ 中的不同维度的判断将其组成的整个空间划分成不同的子空间 (每个空间都代表不同的决策结果), 如图8-10所示。

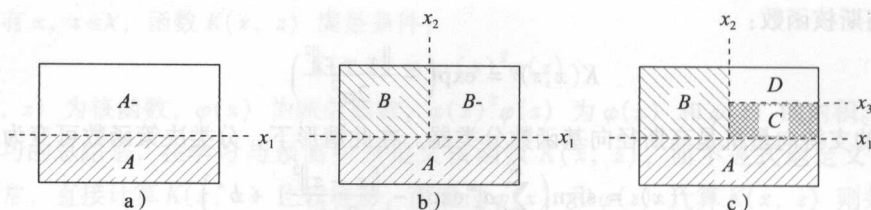


图 8-10 空间划分

假设整个空间用一个矩形来表示，上述决策的第一步（通过对 x_1 的判断）将矩形分为 A 和 A^- 两个部分（ A 部分表示“不约”）；第二步（通过对 x_2 的判断）将 A^- 分为 B 和 B^- 两个部分（ B 部分表示“不约”）；第三步（通过对 x_3 的判断）将 B^- 分为 C 和 D 两个部分（ C 部分表示“不约”、 D 部分表示“约会”）。

上述例子是用决策树描述一个常见的决策过程，而在机器学习场景下，如何根据给定的训练集，自动发现（学习）出这样的决策树，这个过程就是决策树学习。一般的决策树学习的流程大致如下：

输入：训练集

输出：决策树

Step1：寻找合适的属性作为根节点。

Step2：属性的每个可能的值产生一个分支。

Step3：将训练样本划分到适当的分支形成儿子节点。

Step4：重复上面的过程，直到所有的节点都是叶子节点为止。

上述流程第 1 步中的所谓“合适的属性”指的是分类能力最好的属性（借此可以保证训练的决策树深度尽可能小），如何度量每个属性的分类能力是决策树学习中最为重要的问题。基于不同的度量方法和策略，衍生出不同的决策算法，典型的有 ID3、C4.5、C5.0、CART 等，其中前三者都是由罗斯·昆兰（Ross Quinlan）分别在 1986 年、1993 年和 1998 年提出的；而 CART（Classification and Regression Trees）则是由利奥·布雷曼（L. Breiman）、杰罗姆·弗里德曼（J. Friedman）、理查德·欧森（R. Olshen）和查尔斯·斯通（C. Stone）在 1984 年提出的。

ID3 在评估每个属性的重要性时使用的度量方法是信息增益，C4.5 在 ID3 的基础上使用信息增益率作为度量的方法（当然也有其他的改进，下面会具体介绍），而 C5.0 是在决策树的学习过程中集成了 Boosting 算法框架（一种集成思路，后文会具体介绍）以达到决策树学习速度更快、内容使用更加高效及生成的决策树更小的目的。下面重点介绍 ID3 和 C4.5，相关的几个基本定义有：

（1）信息熵

信息的基本作用就是消除人们对事物的不确定性，而信息熵就是对信息的一种量化度量

方法（度量的是信息的有序化程度）。1948年，香农提出了“信息熵”的概念，解决了对系统信息的量化度量问题，信息熵公式定义如下：

$$\text{Entropy}(S) = - \sum_{i=1}^C p_i \log_2(p_i)$$

其中： S 表示样本集（在决策树学习中，就是训练集）， C 表示样本集合中类别的个数， p_i 表示第 i 个类的概率，可由类别 i 中含有的样本个数除以总样本数来得到，因此

$$\sum_{i=1}^C p_i = 1$$

一个系统越是有序，信息熵就越低；反之，一个系统越乱，信息熵就越高。

(2) 信息增益

信息增益指期望信息或信息熵的有效减少量，假设样本集 S 按离散属性 F 的 V 个不同取值划分为 $S_v (v=1, 2, \dots, V)$ ，则信息增益定义为：

$$\text{Gain}(S, F) = \text{Entropy}(S) - \text{ExpectedEntropy}(S_F)$$

其中， $\text{ExpectedEntropy}(S_F)$ 表示 $\text{Entropy}(S_F)$ 的（数学）期望，根据数学期望的定义可以得到：

$$\text{ExpectedEntropy}(S_F) = \sum_{v=1}^V P_v \text{Entropy}(S_v) = - \sum_{v=1}^V P_v \sum_{j=1}^C P_{vj} \log_2(P_{vj})$$

其中， P_{vj} 表示 S_v 中第 j 类的概率，代入上面信息增益的公式，可以得到：

$$\text{Gain}(S, F) = - \sum_{j=1}^C P_j \log_2(P_j) + \sum_{v=1}^V P_v \sum_{j=1}^C P_{vj} \log_2(P_{vj})$$

决策树算法ID3引入了信息增益作为属性选择的标准，并且将建树的方法嵌入其中，其核心是在决策树的各级节点上选择属性时，用信息增益作为属性选择的标准。ID3算法流程如下所述：

输入：训练集 DataSet，属性集 featureList

输出：决策树

Step1：创建根节点 R。

Step2：如果当前 DataSet 中的数据都属于同一类，则标记 R 的类别为该类。

Step3：如果当前 featureList 集合为空，则标记 R 的类别为当前 DataSet 中样本最多的类别。

Step4：递归

Step4-1：从 featureList 中选择属性 F（选择 $\text{Gain}(\text{DataSet}, F)$ 中最大的属性）。

Step4-2：根据 F 的每一个值 v，将 DataSet 划分为不同的子集 DS，对于每一个 DS：

1) 创建节点 C。

2) 如果 DS 为空，则将节点 C 标记为 DataSet 中样本最多的类别。

3) 如果 DS 不为空，则节点 $C = \text{ID3}(\text{DS}, \text{featureList}-F)$ 。

4) 将节点 C 添加为 R 的子节点。

为了便于理解上述流程,下面以一个具体的示例解释上述的计算流程,训练集如表 8-6 所示。

表 8-6 训练集示意表

Day	Outlook	Temperature	Humidity	Wind	Play ball
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

由训练集可知,4 个属性及其值域分别是 Outlook = {Sunny, Overcast, Rain}、Temperature = {Hot, Mild, Cool}、Humidity = {High, Normal}、Wind = {Weak, Strong}; 目标属性及其值域是 Play ball = {Yes, No}。

从数据集中可以看到训练集的样本总数是 14 个,其中类别为 Yes 的个数有 9 个,类别为 No 的有 5 个,令 $P_1 = \frac{9}{14}$ 、 $P_2 = \frac{5}{14}$, 则训练集的信息熵是:

$$\text{Entropy}(S) = - \sum_{i=1}^c p_i \log_2(p_i) = - \frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940\ 286$$

下面比较四个属性 Outlook、Temperature、Humidity、Wind 以选择最有分类能力的属性。先考虑属性 Outlook,该属性有 3 个取值可能 (Sunny, Overcast, Rain), 因此可以将上述训练集划分成 3 个,即 S_1 (Outlook 是 Sunny 的样本,记为 $S_{\text{Outlook}=\text{"Sunny"}}$, 对应训练集中的 D1、D2、D8、D9、D11)、 S_2 (Outlook 是 Overcast 的样本,记为 $S_{\text{Outlook}=\text{"Overcast"}}$, 对应训练集中的 D3、D7、D12、D13) 和 S_3 (Outlook 是 Rain 的样本,记为 $S_{\text{Outlook}=\text{"Rain"}}$, 对应训练集中的 D4、D5、D6、D10、D14)。下面考察样本集 $S_{\text{Outlook}=\text{"Sunny"}}$, 即 S_1 。 S_1 的样本总数为 5, 其中类别为 Yes 的有 2 个 (D9、D11), 类别为 No 的有 3 个 (D1、D2、D8), 记为:

$$\text{Entropy}(S_{\text{Outlook}=\text{"Sunny"}}) = - \sum_{i=1}^c p_i \log_2(p_i) = - \frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.970\ 951$$

特别注意到 $S_{\text{Outlook}=\text{"Sunny"}}$ 的个数有 5 个, 而整个训练集有 14 个样本, 即 $S_{\text{Outlook}=\text{"Sunny"}}$ 占整个训练集

的概率 $P_1 = \frac{5}{14}$, 因此类推可以分别计算 $S_{\text{Outlook} = \text{"Overcast"}}$ 和 $S_{\text{Outlook} = \text{"Rain"}}$ 的信息熵及概率分别是:

$$\text{Entropy}(S_{\text{Outlook} = \text{"Overcast"}}) = - \sum_{i=1}^c p_i \log_2(p_i) = - \frac{4}{4} \log_2 \frac{4}{4} = 0$$

特别注意到, 样本集 $S_{\text{Outlook} = \text{"Overcast"}}$ 中只有一个类别属性 ("Yes"), $S_{\text{Outlook} = \text{"Overcast"}}$ 占整个训练集的概率 $P_2 = \frac{4}{14}$ 。

$$\text{Entropy}(S_{\text{Outlook} = \text{"Rain"}}) = - \sum_{i=1}^c p_i \log_2(p_i) = - \frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.970\ 951$$

此外, $S_{\text{Outlook} = \text{"Rain"}}$ 占整个训练集的概率 $P_3 = \frac{5}{14}$, (属性 Outlook 的) 信息增益率是:

$$\text{Gain}(S, \text{Outlook}) = 0.940\ 286 - \frac{5}{14} \times 0.970\ 51 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.970\ 951 = 0.246\ 75$$

以类似的算法, 可以计算出: $\text{Gain}(S, \text{Temperature}) = 0.029$; $\text{Gain}(S, \text{Humidity}) = 0.151$; $\text{Gain}(S, \text{Wind}) = 0.048$, 由于 $\text{Gain}(S, \text{Outlook})$ 最大, 因此选择根节点划分属性为 Outlook, 则原始的训练集根据 Outlook 的取值可划分为 3 个部分, 即 $S_{\text{Outlook} = \text{"Sunny"}}$ 、 $S_{\text{Outlook} = \text{"Overcast"}}$ 和 $S_{\text{Outlook} = \text{"Rain"}}$, 如图 8-11 所示。

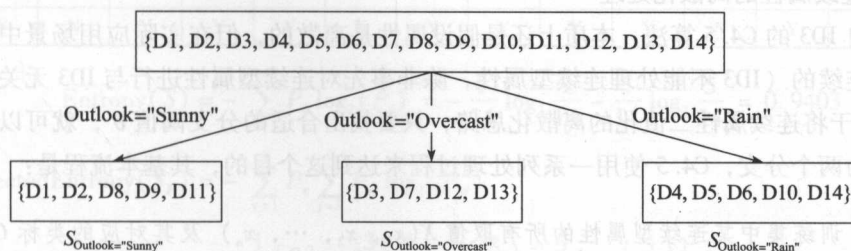


图 8-11 示例训练集第一次划分 (按 Outlook 属性)

由于样本集 $S_{\text{Outlook} = \text{"Overcast"}}$ 中只有一个类标, 这就意味着, 此样本集已经不需要继续划分了 (即此节点已经是叶子节点); 而对于另外的两个划分集 $S_{\text{Outlook} = \text{"Sunny"}}$ 和 $S_{\text{Outlook} = \text{"Rain"}}$ 分别采用上述的方案选择属性继续进行划分, 此处不再赘述。

5. C4.5

C4.5 是昆兰在 1993 年对 ID3 进行的改进算法, 具体动机和思路主要集中在:

1) ID3 算法采用信息增益进行样本集的划分, 信息增益的缺点是倾向于选择取值较多的属性, 在有些情况下这类属性可能不会提供太多有价值的信息。为了解决这个问题, C4.5 算法在做属性选择的时候, 采用信息增益率作为度量依据。

2) ID3 算法无法处理连续值的属性, C4.5 算法既可以处理离散型描述属性, 也可以处理连续型描述属性, 在选择某节点上的分枝属性时, 对于离散型描述属性, C4.5 算法的处理方法与 ID3 相同; 对于连续型属性, C4.5 给出了一个有效的连续属性离散化方案。

3) ID3 算法无法处理属性缺失的情况, C4.5 给出了一个有效的缺失属性的处理方案。

4) C4.5 算法使用训练样本来估计剪枝前后的误差以决定是否真正剪枝, 借此避免树的高度无节制增长, 从而避免数据过度拟合。

以下逐一介绍上述改进动机和思路的具体细节。

(1) 信息增益率

假设样本集 S 按离散属性 F 的 V 个不同取值划分为 $S_v (v=1, 2, \dots, V)$, 则信息增益率定义为:

$$\text{GainRatio}(S, F) = \frac{\text{Gain}(S, F)}{\text{Split}(S, F)}$$

其中, $\text{Split}(S, F)$ 被称为分裂信息, 代表了按照属性 F 分裂样本集 S 的广度和均匀性, 定义如下:

$$\text{Split}(S, F) = - \sum_{v=1}^V \frac{|S_v|}{|S|} \log_2 \left(\frac{|S_v|}{|S|} \right)$$

其中, $|S_v|$ 表示样本集 S_v 中的样本个数, $|S|$ 表示样本集 S 中的样本个数, 因此, 信息增益率中引入 $\text{Split}(S, F)$ 相当于增加了一个惩罚因子 (以避免选择取值过多的属性)。

(2) 连续属性的离散化处理

继承自 ID3 的 C4.5 算法, 本质上还是假设属性是离散的, 但在实际应用场景中, 很多属性往往是连续的 (ID3 不能处理连续型属性, 除非事先对连续型属性进行与 ID3 无关的离散化处理)。基于将连续属性二值化的离散化思路, 只要找出合适的分支阈值 θ' , 就可以将连续的属性划分为两个分支, C4.5 使用一系列处理过程来达到这个目的, 其基本流程是:

输入: 训练集中某连续型属性的所有取值 $X(x_1, x_2, \dots, x_n)$ 及其对应的类标 $C(c_1, c_2, \dots, c_n)$

输出: 线性属性离散化的阈值 θ'

Step1: 将 (x_1, x_2, \dots, x_n) 从小到大进行排序。

Step2: 生成候选阈值 $\theta_i = \frac{x_i + x_{i+1}}{2} (i=1, 2, \dots, n-1)$ 。

Step3: 用信息增益率选择最佳划分 θ' 。

为了便于理解上述流程, 以一个具体的示例加以解释, 假设某个连续属性的取值及对应的类标如表 8-7 所示, 类别属性包括 “Y” 和 “N” 两类, 为了表示方便, 已经将连续属性按大小排序 (共 14 个)。

表 8-7 示例样本

X	64	65	68	69	70	71	72	72	75	75	80	81	83	85
C	Y	N	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y	N

根据 $\theta_i = \frac{x_i + x_{i+1}}{2}$ ($i = 1, 2, \dots, n-1$) 生成候选阈值 (共 13 个), 参见表 8-8。

表 8-8 候选阈值

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	θ_{13}
64.5	66.5	68.5	69.5	70.5	71.5	72.0	73.5	75.0	77.5	80.5	82.0	84.0

由前面推导可知信息增益率计算公式如下:

$$\text{GainRatio}(S, F) = \frac{\text{Gain}(S, F)}{\text{Split}(S, F)} = \frac{\text{Entropy}(S) - \text{ExpectedEntropy}(S_F)}{\text{Split}(S, F)}$$

$$\text{GainRatio}(S, F) = \frac{\text{Gain}(S, F)}{\text{Split}(S, F)} = \frac{-\sum_{j=1}^C P_j \log_2(P_j) + \sum_{v=1}^V P_v \sum_{j=1}^C P_{vj} \log_2(P_{vj})}{-\sum_{v=1}^V \frac{|S_v|}{|S|} \log_2\left(\frac{|S_v|}{|S|}\right)}$$

如果阈值取 θ_1 , 则训练样本被分为两个分支 (S_1 和 S_2), 如表 8-9 所示。

表 8-9 样本分割

	S_1	S_2											
X	64	65	68	69	70	71	72	72	75	75	80	81	83
C	Y	N	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y

$$\text{Entropy}(S) = -\sum_{j=1}^C P_j \log_2(P_j) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.9403$$

$$\text{ExpectedEntropy}(S_F) = -\sum_{v=1}^V P_v \sum_{j=1}^C P_{vj} \log_2(P_{vj})$$

$$= -\frac{1}{14}[0] - \frac{13}{14}\left[\frac{8}{13} \log_2\left(\frac{8}{13}\right) + \frac{5}{13} \log_2\left(\frac{5}{13}\right)\right] = 0.8926$$

$$\text{Split}(S, F) = -\sum_{v=1}^V \frac{|S_v|}{|S|} \log_2\left(\frac{|S_v|}{|S|}\right) = -\frac{1}{14} \log_2 \frac{1}{14} - \frac{13}{14} \log_2 \frac{13}{14} = 0.3712$$

$$\text{所以 } \text{GainRatio}(S, F)_{\theta_1} = \frac{\text{Entropy}(S) - \text{ExpectedEntropy}(S_F)}{\text{Split}(S, F)} = 0.1285。$$

以此类推, 可以分别求得 θ_i ($i = 2, 3, \dots, n-1$) 情况下的 $\text{GainRatio}(S, F)_{\theta_i}$, 结果罗列如表 8-10 所示。

表 8-10 不同 θ_i 下的 $\text{GainRatio}(S, F)_{\theta_i}$

i	θ_i	Entropy(S)	ExpectedEntropy(S_F)	Split(S, F)	GainRatio(S, F)
1	64.5	0.9403	0.8926	0.3712	0.1285
2	66.5	0.9403	0.9300	0.5917	0.0175
3	68.5	0.9403	0.9398	0.7496	0.0007
4	69.5	0.9403	0.9253	0.8631	0.0173
5	70.5	0.9403	0.8950	0.9403	0.0482

(续)

i	θ_i	Entropy (S)	ExpectedEntropy (S_F)	Split (S, F)	GainRatio (S, F)
6	71.5	0.9403	0.9389	0.9852	0.0014
7	72.0	0.9403	0.9389	0.9852	0.0014
8	73.5	0.9403	0.9389	0.9852	0.0014
9	75.0	0.9403	0.9152	0.8631	0.0291
10	77.5	0.9403	0.9152	0.8631	0.0291
11	80.5	0.9403	0.9398	0.7496	0.0007
12	82.0	0.9403	0.9300	0.5917	0.0175
13	84.0	0.9403	0.8269	0.3712	0.3055

由表 8-10 可以看到，当 $\theta = 84.0$ 时，GainRatio(S, F) 最大，因此，选择的阈值应该是 84.0。

从上面的推导过程中可以发现，C4.5 的连续属性离散化的处理策略存在如下几个缺陷及相应的解决方案：

1) 如果连续属性有 N 种取值可能，为了找到合适的阈值，需要做 $N-1$ 次计算，重复计算量很大。一个改进思路是：对连续属性进行排序，在分类结果产生变化的两组数据之间取分割点，这样，可以有效地减少计算量。以上述示例为例，需要考虑的阈值见表 8-11 所示。

表 8-11 分割点的优化选择

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	θ_{13}
64.5	66.5	—	—	70.5	—	72.0	—	—	77.5	80.5	—	84.0

2) 缺陷及改进思路：基于信息增益率的连续属性离散化方法，分割点的位置会影响到信息增益率的值。当趋于平均分割时，增益率的抑制达到最大化，子集样本的个数能够影响分界点，这显然是不合理的。一个改进思路是：确定最优分割点的原则就是要求该处分割达到信息增益的最大化。而在选择分裂属性时，仍使用增益率。

(3) 缺失属性处理

所谓缺失属性，指的是训练样本（或者测试样本）中的某条实例 $X(x_1, x_2, \dots, x_n)$ 的某一个（或者多个）属性值 $x_i(i \in \{0, 1, 2, \dots, n\})$ （记为 a ）为空。在 C4.5 算法中，根据缺失属性出现的不同场景，可采用不同的处理策略，具体见表 8-12。

表 8-12 缺失属性处理策略

场景	可选策略
1 利用信息增益或信息增益率进行属性的选择（进行分支）时，当前考察的属性具有缺失属性	[1-1] 忽略具有缺失属性的训练样本实例 [1-2] 赋予缺失属性一个均值或最常用的值（考察所有其他样本实例该属性的值） [1-3] 计算增益或增益率时根据缺失属性样本的个数所占的比率对增益/增益率进行相应的“打折” [1-4] 根据其他属性利用其他算法（需要重新设计）把这些样本缺失的属性补全

(续)

	场景	可选策略
2	一个属性已被选择的情况下, 进行样本划分的时候, 有些样本存在缺失属性	<p>[2-1] 忽略具有缺失属性的训练样本实例</p> <p>[2-2] 赋予缺失属性一个均值或最常用的值 (考察所有其他样本实例该属性的值)</p> <p>[2-3] 根据其他属性利用其他算法 (需要重新设计) 把这些样本缺失的属性补全, 然后继续处理将其划分到相应的子集</p> <p>[2-4] 单独为缺失属性的样本划分一个分支子集</p> <p>[2-5] 把这些缺失属性的样本, 按照具有属性 a 的样本被划分成的子集样本个数的相对比率, 分配到各个子集中去 (至于如何分类, 需要重新设计方法或策略)</p> <p>[2-6] 把缺失属性的样本分配给所有的子集, 也就是说每个子集都有这些缺失属性的样本</p>
3	决策树已经训练生成, 但待分类的测试样本缺失了某些属性	<p>[3-1] 如果有单独的缺失分支 (利用上述策略 2-4), 则依据此分支</p> <p>[3-2] 把待分类的样本的属性 a 值分配一个均值或最常用的值 (考察所有其他样本实例该属性的值), 然后进行分支预测</p> <p>[3-3] 根据其他属性利用其他算法 (需要重新设计) 把这些样本缺失的属性补全, 然后进行分支处理</p> <p>[3-4] 在决策树中属性 a 节点的分支上, 遍历属性 a 节点的所有分支, 探索所有可能的分类结果, 然后把这些分类结果结合起来一起考虑, 按照概率决定一个分类</p> <p>[3-5] 待分类样本在到达属性 a 节点时就终止分类, 然后根据此时 a 节点所覆盖的叶子节点类别状况为其分配一个发生概率最高的类</p>

C4.5 算法流程如下:

输入: 训练集 DataSet; 属性集 featureList

输出: 决策树

Step1: 创建根节点 R。

Step2: 如果当前 DataSet 中的数据都属于同一类, 则标记 R 的类别为该类。

Step3: 如果当前 featureList 的集合为空, 则标记 R 的类别为当前 DataSet 中样本最多的类别。

Step4: 递归

Step4-1: 从 featureList 中选择属性 F (选择 GainRatio (DataSet, F) 最大的属性, 连续属性参见上面的离散化过程)。

Step4-2: 根据 F 的每一个值 v , 将 DataSet 划分为不同的子集 DS, 对于每一个 DS:

- 1) 创建节点 C。
- 2) 如果 DS 为空, 则将节点 C 标记为 DataSet 中样本最多的类别。
- 3) 如果 DS 不为空, 则节点 $C = C4.5(DS, \text{featureList}-F)$ 。
- 4) 将节点 C 添加为 R 的子节点。

6. CART

CART(Classification And Regression Tree, 分类回归树)是由利奥·布雷曼、杰罗姆·弗里德曼、理查德·欧森和查尔斯·斯通在 1984 年提出的一种既可以用于分类 (输出是样本的类

标),也可以用于回归(输出是实数)的决策树。CART的几个典型特征包括:

1) 二元划分:由于二叉树不易产生数据碎片,精确度往往也会高于多叉树,所以在CART算法中,采用了二元划分,即CART是一棵二叉树,且每个非叶子节点都有两个孩子。

2) 不纯度度量:对于离散型属性,CART中用于选择属性的不纯度度量是Gini指数,如果目标变量是连续的,则CART算法找出一组基于树的回归方程来预测目标变量。

3) 剪枝策略:CART用独立的验证数据集对训练集生长的树进行剪枝。

Gini指数是用来度量数据集不纯性的方法(作用类似前文提到的信息增益和信息增益率),其定义如下:

$$\text{Gini}(S) = 1 - \sum_{i=1}^c p_i^2$$

其中, S 是指训练集, p_i 是指 S 中类别 i 的概率(假设该训练集中共有 C 个类别),由此可见, $\text{Gini}(S)$ 越小,则表明训练集 S 越纯。

CART中考虑每个属性上的二元划分(最小化Gini指数),根据训练集 S 中的属性 F 分为 S_1 和 S_2 ,则给定划分的Gini指数定义如下:

$$\text{Gini}_F(S) = \frac{|S_1|}{|S|} \text{Gini}(S_1) + \frac{|S_2|}{|S|} \text{Gini}(S_2)$$

为了便于理解Gini指数的计算,下面以一个具体的示例解释上述的计算流程,训练集如表8-13所示(上文实际上使用过类似的例子)。

表 8-13 训练集示意图

Day	Outlook	Temperature	Humidity	Wind	Play ball
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

作为示例,仅考察离散属性 $\text{Outlook} = \{\text{Sunny}, \text{Overcast}, \text{Rain}\}$,由于Outlook有3个属性,

因此从二元划分的角度出发, 仅有3种划分方案:

1) $F_1: \{\text{Sunny}\}, \{\text{Overcast}, \text{Rain}\}$, 此种分配方案下, 训练集被分成了2个部分, 其中 $S_1 = \{D1, D2, D8, D9, D11\}$, $S_2 = \{D3, D4, D5, D6, D7, D10, D12, D13, D14\}$, 在训练集 S_1 中, 类标为“**Yes**”的有2个 ($D9, D11$), 类标为“**No**”的有3个 ($D1, D2, D8$); 在训练集 S_2 中, 类标为“**Yes**”的有7个 ($D3, D4, D5, D7, D10, D12, D13$), 类标为“**No**”的有2个 ($D6, D14$)。因此

$$\text{Gini}(S_1) = 1 - \left(\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right) = 0.48$$

$$\text{Gini}(S_2) = 1 - \left(\left(\frac{7}{9} \right)^2 + \left(\frac{2}{9} \right)^2 \right) = 0.3457$$

$$\text{Gini}_{F_1}(S) = \frac{5}{14}\text{Gini}(S_1) + \frac{9}{14}\text{Gini}(S_2) = 0.3937$$

2) $F_2: \{\text{Rain}\}, \{\text{Sunny}, \text{Overcast}\}$

同上, 可求得 $\text{Gini}_{F_2}(S) = 0.4571$ 。

3) $F_3: \{\text{Overcast}\}, \{\text{Sunny}, \text{Rain}\}$

同上, 可求得 $\text{Gini}_{F_3}(S) = 0.3571$ 。

由于 F_3 划分下, Gini 指数最小, 因此应当以 F_3 进行划分。

分裂属性的选择规则是: 选择具有最小 Gini 指数的属性为分裂属性, 这个规则需要对每个属性都要遍历所有可能的分割方法。对于离散值属性, 在算法中递归地选择该属性产生最小 Gini 指数的子集作为它的分裂子集。对于连续值属性, 必须考虑所有可能的划分点, 其策略类似于 C4.5 中介绍的方法, 公式如下:

$$\text{Gini}_A(D) = \sum_{i=1}^v \frac{|D_i|}{|D|} \text{Gini}(D_i)$$

其中, A 是待考察的连续型属性, v 是可能的划分点的个数。

CART 算法满足下列条件之一的, 即可视为叶节点不再进行分支操作:

1) 所有叶节点的样本数为1, 或者样本数小于某个给定的最小值, 或者样本都属于同一类的时候。

2) 决策树的高度达到用户设置的阈值, 或者分支后的叶节点中的样本属性都属于同一个类的时候。

3) 当训练数据集中不再有属性向量作为分支选择的时候。

CART 分类树的算法流程如下:

输入: 训练集 DataSet、属性列表 featureList、阈值 alpha (用于控制决策树深度)

输出: CART 分类树

Step1: 创建根节点 R。

Step2: 如果当前 DataSet 中数据的类别相同, 则标记 R 的类别为该类。

Step3: 如果决策树的高度大于 α , 则不再分解, 标记 R 的类别为 `classify (DataSet)`。

Step4: 递归

Step4-1: 标记 R 的类别 `classify (DataSet)`。

Step4-2: 从 `featureList` 中选择属性 F (选择 `Gini (DataSet, F)` 最小的属性划分, 连续属性参考 C4.5 的离散化过程 (以 Gini 最小作为划分标准))。

Step4-3: 根据 F, 将 `DataSet` 做二元划分 `DS_L` 和 `DS_R`, 如果 `DS_L` 或 `DS_R` 为空, 则不再分解, 如果 `DS_L` 和 `DS_R` 都不为空, 则:

1) $C_L = \text{CART_classification}(DS_L, \text{featureList}, \alpha)$ 。

2) $C_R = \text{CART_classification}(DS_R, \text{featureList}, \alpha)$ 。

3) 将节点 C_L 和 C_R 添加为 R 的左右子节点。

CART 除了可以用于如上所述的分类之外, 还可以用于回归。与分类的场景不同的是, 用于回归的训练集 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 $y_i (i=1, 2, \dots, N)$ 是每个训练样本的因变量 (其取值为实数), N 为训练集样本的个数。

设 t 代表树的某个节点, 该节点中的样本集合为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_{N_t}, y_{N_t})\}$, 其中 N_t 是节点 t 中样本的个数。节点 t 的因变量的均值为:

$$\bar{y}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} y_i$$

节点 t 内的平方残差最小化 (Squared Residuals Minimization Algorithm) 定义为:

$$\text{SRMA}_t = \sum_{i=1}^{N_t} (y_i - \bar{y}_t)^2$$

划分 (属性) F 将 t 划分成左右节点 t_L 和 t_R , 则 ϕ 值定义为:

$$\phi(t, F) = \text{SRMA}_t - \text{SRMA}_{t_L} - \text{SRMA}_{t_R}$$

使上式取值最大的划分 (属性) F^* 是最佳的属性划分, 即

$$\phi(t, F^*) = \max(\phi(t, F))$$

CART 回归树的算法流程如下:

输入: 训练集 `DataSet`、属性集 `featureList`、阈值 α (控制深度)、阈值 δ (控制分裂)

输出: CART 回归树

Step1: 创建根节点 R。

Step2: 如果当前 `DataSet` 中数据的值都相同, 则标记 R 的值为该值。

Step3: 如果最大的 ϕ 值小于设定阈值 δ , 则标记 R 的值为 `DataSet` 因变量均值。

Step4: 如果其中一个要产生的节点的样本数量小于 α , 则不再分解, 标记 R 的值为 `DataSet` 因变量均值。

Step5: 递归

Step5-1: 从 featureList 中选择属性 F, 选择 $\phi(\text{DataSet}, F)$ 最大的属性, 连续属性 (或使用多个属性的线性组合) 参考 C4.5 的离散化过程 (以 ϕ 最大作为划分标准)。

Step5-2: 根据 F, 将 DataSet 做二元划分 DS_L 和 DS_R, 如果 DS_L 或 DS_R 为空, 则标记节点 R 的值为 DataSet 因变量均值, 如果 DS_L 和 DS_R 都不为空, 则:

1) $C_L = \text{CART_regression}(\text{DS_L}, \text{featureList}, \alpha, \delta)$ 。

2) $C_R = \text{CART_regression}(\text{DS_R}, \text{featureList}, \alpha, \delta)$ 。

3) 将节点 C_L 和 C_R 添加为 R 的左右子节点。

7. 随机森林

随机森林 (Random Forest) 是将 Bagging 算法用于多个 CART 决策树的构建, 并在多个 CART 决策树中采用简单多数投票法 (针对分类) 或单棵树输出结果的简单平均 (针对回归) 以得到最终结果。

Bagging 算法, 名称源于 bootstrap aggregation (自助聚集), 是一种用来提高学习算法准确度的集成方法, 这种方法通过构造一个预测函数系列, 然后以一定的方式将它们组合成一个预测函数。Bagging 的流程如图 8-12 所示, Bagging 的基本流程是: 从大小为 N 的原始数据集 D 中, 分别独立随机地抽取 N' ($N' < N$) 个数据形成新的数据集 D_i ($i=1, 2, \dots, B$), 每个独立的数据集独立训练生成 B 个独立的基分类器 T_i ($i=1, 2, \dots, B$)。最后的分类判别将根据这些独立的分类器各自的判决结果的简单投票来决定。

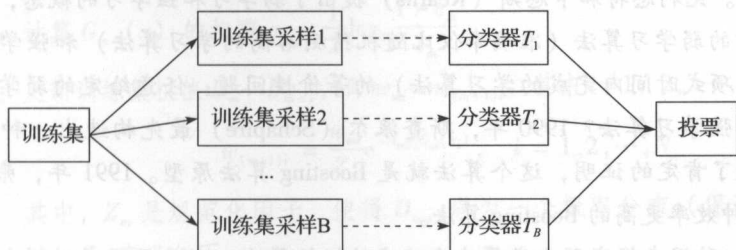


图 8-12 Bagging 流程

值得注意的是:

1) 通常这些分量分类器的模型都是一样的, 如神经网络、SVM、决策树等, 并且最好这些模型都是非稳定的, 即: 训练集的改变对模型的影响比较大。

2) Bagging 也可用于回归, 将上述的简单投票改为 (加权) 平均即可。

随机森林的算法要点有:

1) 使用 Bagging 方法形成每棵树的训练集 (从训练集中随机抽取)。

2) 假设共有 M 个属性, 对于每个内部节点 (包括根节点), 从 M 个属性中随机抽取 F ($F \leq M$) 个属性作分裂属性集, 以这 F 个属性上最好的分裂方式对节点进行分裂 (在整个森林的生长过程中, F 的值一般维持不变)。

3) 每棵树任其生长, 不进行剪枝。

两个随机性的引入 (从训练集中随机抽取子训练集及随机抽取属性), 使得随机森林不会轻易陷入过拟合并且具有较好的抗噪声能力, 同时, 由于随机森林是在 CART 的基础上进行的 Bagging, 因此其既能处理离散型数据, 也能处理连续型数据, 数据集无须规范化, 影响随机森林分类性能的主要因素有:

1) 森林中单棵树的分类强度 (Strength): 每棵树的分类强度越大, 则随机森林的分类性能越好。

2) 森林中树与树之间的相关度 (Correlation): 树与树之间的相关度越大, 则随机森林的分类性能越差。

8. AdaBoost

AdaBoost (Adaptive Boosting, 意为“自适应增强”, 一般不翻译) 是由约阿夫·弗洛因德 (Yoav Freund) 和罗伯特·斯查派尔 (Robert Schapire) 在 1995 年提出的一种基于集成思想的框架算法 (迭代算法), 该框架算法的核心思想是针对同一个训练集训练不同的分类器 (弱分类器, 准确率仅比随机猜测略高的分类器算法), 然后把这些弱分类器集合起来, 构成一个更强的最终分类器 (强分类器, 准确率很高并能在多项式时间内完成的学习算法)。

Boosting 的思想起源于瓦利恩特 (Valiant) 提出的 PAC (Probably Approximately Correct, PAC) 学习模型。瓦利恩特和卡恩斯 (Kearns) 提出了弱学习和强学习的概念, 并首次提出了 PAC 学习模型中的弱学习算法 (准确率仅比随机猜测略高的学习算法) 和强学习算法 (准确率很高并能在多项式时间内完成的学习算法) 的等价性问题: 任意给定的弱学习算法, 是否可以将其提升为强学习算法? 1990 年, 斯查派尔 (Schapire) 最先构造出一种多项式级的算法, 对该问题做了肯定的证明, 这个算法就是 Boosting 算法原型。1991 年, 弗洛因德 (Freund) 提出了一种效率更高的 Boosting 算法。

Boosting 是一种用来提高弱分类算法准确度的框架算法, 主要是通过对样本集的操作获得样本子集, 然后用弱分类算法在样本子集上训练生成一系列的基分类器, 然后 Boosting 框架算法将这些基分类器进行加权融合, 产生一个最后的结果分类器。在这些基分类器中, 每个单个的分类器的识别率不一定很高, 但他们联合后的结果有很高的识别率, 这样便提高了该弱分类算法的识别率。在产生单个的基分类器时, 可用相同的分类算法, 也可用不同的分类算法 (这些算法一般是不稳定的弱分类算法, 如神经网络、决策树等, 所谓不稳定指的是数据集中小的变动都能够使得分类结果发生显著的变动)。

但是, 这两种算法存在共同的实践上的缺陷: 需要事先知道弱学习算法学习正确率的下限。1995 年, Freund 和 Schapire 在 Boosting 算法的基础上提出了 AdaBoost (Adaptive Boosting)

算法, 该算法的效率和 Freund 于 1991 年提出的 Boosting 算法几乎相同, 但不需要任何关于弱学习器的先验知识, 因而更容易应用到实际问题当中。

AdaBoost 是 Boosting 算法家族中的一个代表算法, 该算法主要是在整个训练集上维护一个权重向量 $D_i (i=1, 2, \dots, N)$ (其中 N 表示训练集中的训练样本的个数), 用赋予权重的训练集通过弱分类算法产生基分类器, 通过该基分类器的误差率更新训练样本的权重向量 $D_i (i=1, 2, \dots, N)$ (对错误分类的样本分配更大的权值, 对正确分类的样本赋予更小的权值)。每次更新后用相同的弱分类算法产生新的基分类器, 这些基分类器构成多分类器, 然后对这些多分类器用加权的方法进行集成, 最后得到决策结果。AdaBoost 算法流程如下:

输入: 训练集 $\{(x_1, y_1), (x_1, y_2), \dots, (x_N, y_N)\}$

输出: 联合的分类器及权值分布

Step1: 初始化训练数据的样本权重分布, 每一个训练样本都被赋予相同的权重 $\frac{1}{N}$, 即

$$D_m = (w_{m1}, w_{m2}, \dots, w_{mi}, \dots, w_{mN}), \quad w_{mi} = \frac{1}{N}, \quad i = 1, 2, \dots, N; \quad m = 1, 2, \dots, M$$

其中, D_m 表示第 m 次迭代时的样本权重分布 (假设进行 M 轮迭代), w_{mi} 表示第 m 次迭代时第 i 条样本的权重 (假设训练集共有 N 条训练样本), 程序开始时 $m=1$ 。

Step2: 对于 $m=1, 2, \dots, M$

Step2-1: 使用具有权值分布 D_m 的训练集学习, 得到基分类器 $G_m(x)$ 。

Step2-2: 计算 $G_m(x)$ 在训练集上的分类误差率 $e_m = P(G_m(x_i) \neq y_i)$ 。

Step2-3: 计算 $G_m(x)$ 的权重 $a_m = \frac{1}{2} \log_2 \frac{1-e_m}{e_m}$ 。

Step2-4: 更新训练集权值 $D_m \rightarrow D_{m+1}$, 即 $w_{mi} \rightarrow w_{(m+1)i}$, 其中 $i=1, 2, \dots, N$:

$$w_{(m+1)i} = \frac{w_{mi}}{Z_m} e^{-a_m y_i G_m(x_i)}, \quad i = 1, 2, \dots, N$$

其中, Z_m 是规范化因子, 使得 D_{m+1} 成为一个概率分布 (保证权值和为 1), Z_m 定义如下:

$$Z_m = \sum_{i=1}^N w_{mi} e^{-a_m y_i G_m(x_i)}$$

Step3: 构造基分类器的线性组合 $f(x) = \sum_{m=1}^M a_m G_m(x)$, 则最终分类器是:

$$G(x) = \text{sign}(f(x)) = \text{sign}\left(\sum_{m=1}^M a_m G_m(x)\right)$$

为了更好地理解上述流程, 下面以简单的二值训练集为例介绍具体的实现过程, 使用的数据集如表 8-14 所示。

表 8-14 数据集示意

i	1	2	3	4	5	6	7	8	9	10
x	0	1	2	3	4	5	6	7	8	9
y	1	1	1	-1	-1	-1	1	1	1	-1

如表 8-14 所示, 该训练集共有 10 个训练样本, 即 $N=10$; 该训练集的类标只有两个 (1 和 -1), 求解过程如下:

Step1: 初始化每个权值分布, 如下所示:

$$D_m = (w_{m1}, w_{m2}, \dots, w_{mi}, \dots, w_{m10}), w_{mi} = \frac{1}{10}, \quad i = 1, 2, \dots, 10, m = 1$$

Step2: 对于 $m=1$, 在权值分布为 D_1 的训练数据上, 不同的阈值可以生成不同的分类器, 所有可能的阈值及其对应的误差率如表 8-15 所示。

表 8-15 阈值及其对应的误差率

阈值 ζ	基分类器	错误点	误差率
0.5	$G_1(x) = \begin{cases} 1, & x < \zeta \\ -1, & x > \zeta \end{cases}$	2, 3, 7, 8, 9	$D_{m2} + D_{m3} + D_{m7} + D_{m8} + D_{m9} = \frac{5}{10}$
1.5		3, 7, 8, 9	$D_{m3} + D_{m7} + D_{m8} + D_{m9} = \frac{4}{10}$
2.5		7, 8, 9	$D_{m7} + D_{m8} + D_{m9} = \frac{3}{10}$
3.5		4, 7, 8, 9	$D_{m4} + D_{m7} + D_{m8} + D_{m9} = \frac{4}{10}$
4.5		4, 5, 7, 8, 9	$D_{m4} + D_{m5} + D_{m7} + D_{m8} + D_{m9} = \frac{5}{10}$
5.5		4, 5, 6, 7, 8, 9	$D_{m4} + D_{m5} + D_{m6} + D_{m7} + D_{m8} + D_{m9} = \frac{6}{10}$
6.5		4, 5, 6, 8, 9	$D_{m4} + D_{m5} + D_{m6} + D_{m8} + D_{m9} = \frac{5}{10}$
7.5		4, 5, 6, 9	$D_{m4} + D_{m5} + D_{m6} + D_{m9} = \frac{4}{10}$
8.5		4, 5, 6	$D_{m4} + D_{m5} + D_{m6} = \frac{3}{10}$

由上表可知, 阈值取 2.5 或 8.5 时误差率最低, 本示例取 2.5, 则基本分类器是:

$$G_1(x) = \begin{cases} 1, & x < 2.5 \\ -1, & x > 2.5 \end{cases}$$

从而可得 $G_1(x)$ 在训练集上的误差率 $e_1 = P(G_1(x_i) \neq y_i) = 0.3$, 计算 $G_1(x)$ 的权重系数:

$$a_1 = \frac{1}{2} \log_2 \frac{1 - e_1}{e_1} = 0.4236$$

更新训练样本的权值分布 $D_{(m+1)} = (w_{(m+1)1}, w_{(m+1)2}, \dots, w_{(m+1)i}, \dots, w_{(m+1)N}), i = 1, 2, \dots, N$, 其中:

$$w_{(m+1)i} = \frac{w_{mi}}{Z_m} e^{-a_m y_i G_m(x_i)}, \quad i = 1, 2, \dots, N$$

$$Z_m = \sum_{i=1}^N w_{mi} e^{-a_m y_i G_m(x_i)}$$

为了便于示例说明, 权值更新涉及的中间结果如表 8-16 所示。

表 8-16 权值更新过程

i	w_{mi}	$y_i G_m(x_i)$	$e^{-a_m y_i G_m(x_i)}, i = 1, 2, \dots, N$	$w_{(m+1)i} = \frac{w_{mi}}{Z_m} e^{-a_m y_i G_m(x_i)}, i = 1, 2, \dots, N$
1	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715
2	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715
3	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715
4	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715
5	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715
6	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715
7	$\frac{1}{10}$	-1	$\frac{1}{10} \times e^{-0.4236 \times (-1)} = 1.5275$	0.0166
8	$\frac{1}{10}$	-1	$\frac{1}{10} \times e^{-0.4236 \times (-1)} = 1.5275$	0.0166
9	$\frac{1}{10}$	-1	$\frac{1}{10} \times e^{-0.4236 \times (-1)} = 1.5275$	0.0166
10	$\frac{1}{10}$	1	$e^{-0.4236 \times (1)} = 0.6547$	0.0715

分类函数 $f_1(x) = 0.4236G_1(x)$, 最终得到的分类器 $\text{sign}(f_1(x))$ 在训练集上有 3 个错误分类点 (7, 8, 9), 根据表 8-16, 经过一轮迭代以后, 样本的权重分布是:

$$D_2 = (0.0715, 0.0715, 0.0715, 0.0715, 0.0715, 0.0715, 0.1666, 0.1666, 0.1666, 0.0715)$$

Step3: 与 Step2 的做法相同, 对于 $m=2$, 在权值分布为 D_2 的训练数据上, 阈值取 8.5 时误差最低 ($e_2 = P(G_2(x_i) \neq y_i) = 0.2143$), 则基本分类器是

$$G_2(x) = \begin{cases} 1, & x < 8.5 \\ -1, & x > 8.5 \end{cases}$$

计算 $G_2(x)$ 的权重系数:

$$a_2 = \frac{1}{2} \log_2 \frac{1 - e_2}{e_2} = 0.6496$$

更新训练样本权值分布, 与 Step2 的做法相同, 可得到:

$$D_3 = (0.0455, 0.0455, 0.0455, 0.1667, 0.1667, 0.1667, 0.1060, 0.1060, 0.1060, 0.0455)$$

则分类函数 $f_2(x) = 0.4236G_1(x) + 0.6496G_2(x)$, 最终得到的分类器 $\text{sign}(f_2(x))$ 在训练集上

有 3 个错误分类点。

Step4: 对于 $m=3$, 在权值分布为 D_3 的训练数据上, 阈值取 5.5 时误差率最低, 则基本分类器是:

$$G_3(x) = \begin{cases} 1, & x < 5.5 \\ -1, & x > 5.5 \end{cases}$$

从而可得 $G_1(x)$ 在训练集上的误差率 $e_3 = P(G_3(x_i) \neq y_i) = 0.1820$, 计算 $G_2(x)$ 的权重系数:

$$a_3 = \frac{1}{2} \log_2 \frac{1 - e_3}{e_3} = 0.7514$$

更新训练样本权值分布, 与 Step3 的做法相同, 可以得到:

$$D_4 = (0.125, 0.125, 0.125, 0.102, 0.102, 0.102, 0.065, 0.065, 0.065, 0.125)$$

则分类函数 $f_3(x) = 0.4236G_1(x) + 0.6496G_2(x) + 0.7514G_3(x)$, 最终得到的分类器 $\text{sign}(f_3(x))$ 在训练集上有 0 个错误分类点。

需要说明的是, 针对不同的问题场景, 采用不同的思路, 事实上有很多其他类型的分类算法, 并且新的分类算法也在持续的推出中, 本文无法一一罗列所有的分类算法, 仅就上述的典型算法做个简单介绍。

8.5 本章小结

作为数据建模的重要手段之一, 机器学习是一个重要的、可行的候选技术选型, 本章从机器学习过程中有无导师参与的角度介绍了非监督学习和监督学习, 非监督学习是没有导师参与的机器学习, 某种意义上而言, 非监督学习往往专指聚类; 监督学习是有导师参与的机器学习, 所谓有导师参与指的是在给定的训练集中, 除了有属性数据外, 针对每一行属性数据, 还有导师 (领域专家) 为其做的标注, 这实质上是指在训练集中已经包含了导师 (领域专家) 对属性数据的理解, 监督学习就是按照导师对数据的理解趋势 (意图) 建立 “属性 - 标注” 的关系模型, 根据标注的类别是连续的还是离散的, 又将监督学习分为回归和分类两个类别。本章对常见的非监督学习及监督学习方法进行了简单的介绍, 应该说, 理解所有这些方法是接触数据建模或大数据分析相关工作必须具备的基本技能。

事实上, 在很多应用场景中, 我们需要专家对采样数据进行标注 (期望是监督学习), 出于多种原因, 专家并不能对训练集中的所有数据进行标注, 这就导致在训练集中有一部分数据是标注的 (往往是有限的), 还有一类是未被标注的。如果仅利用标注的数据作为训练集就是典型的监督学习, 势必浪费那些未被标注的数据, 因此能够利用有限标注的数据和大量未标注的样本进行学习就成了一个应用驱动的重要研究领域, 这种场景下的机器学习称为半监督学习 (Semi-supervised Learning)。半监督学习是监督学习与无监督学习相结合的一种学习方法, 它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题, 对于减少标注代价、提高机器学习性能具有非常重大的实际意义。

强化学习是另外一类机器学习方法,本质上,强化学习可以看成是一种有导师监督的机器学习方法,不过由于其研究视角和执行思路与传统的监督学习有很大的差异性,因此往往更愿意将其作为一种独立的机器学习分类。在强化学习模式下,输入数据作为对模型的反馈,不像监督模型那样,输入数据仅仅是作为一种检查模型对错的方式,在强化学习下,输入数据直接反馈到模型,模型必须对此立刻做出调整。常见的应用场景包括动态系统及机器人控制等。常见算法包括 Q-Learning 及时序差分学习 (Temporal Difference Learning)。

对于大数据应用场景下的数据建模问题,最常用的方法或许就是监督学习和非监督学习;在某些专门的垂直应用领域,比如图像识别领域,由于存在大量的非标识的数据和少量的可标识数据,半监督学习会很有效;而强化学习更多地应用于机器人控制及其他需要与环境交互并进行系统控制的场景。

子曰:“工欲善其事,必先利其器”(摘自《论语·卫灵公》);子(又)曰:“君子不器”(摘自《论语·为政》)。

对于大数据分析师的提示或许是:①针对任何一个应用目标和分析需求,应该选择最合适的那个方法(利其器)。②对于大数据分析师本人而言,不仅需要具备建模的技能及素养,还需要对目标应用、现场数据、建模方法等都极其敏感。

本章参考文献

- [1] Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms [C]. Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 161-168.
- [2] Gaddam S R, Phoha V V, Balagani K S. K-means + ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-means Clustering and ID3 Decision Tree Learning Methods [J]. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(3): 345-354.
- [3] Hastie T, Tibshirani R, Friedman J. Unsupervised Learning [M]. New York: Springer, 2009.
- [4] Joachims T. Making Large Scale SVM Learning Practical [R]. Universität Dortmund, 1999.
- [5] Leake D B. Case-based Reasoning [J]. The Knowledge Engineering Review, 1994, 9(01): 61-64.
- [6] Liaw A, Wiener M. Classification and Regression by RandomForest [J]. R News, 2002, 2(3): 18-22.
- [7] McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification [C]. AAAI-98 Workshop on Learning for Text Categorization. 1998, 752: 41-48.
- [8] McCorduck P, Minsky M, Selfridge O G, et al. History of Artificial Intelligence [C]. IJCAI. 1977: 951-954.
- [9] McLachlan G, Krishnan T. The EM Algorithm and Extensions [M]. Manhattan: John Wiley & Sons, 2007.
- [10] Quinlan J R. Induction of Decision Trees [J]. Machine Learning, 1986, 1(1): 81-106.
- [11] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means Clustering with Background Knowledge [C]. ICML. 2001, 1: 577-584.

- [12] Zhang M L, Zhou Z H. ML-KNN: A Lazy Learning Approach to Multi-label Learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048.
- [13] 赖丽如. 基于数据挖掘的乳腺癌术后中医辨治与用药规律研究 [D]. 南京中医药大学, 2015.
- [14] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012.
- [15] 左艇. 基于大数据环境下临床“十八反”同方配伍规律的研究 [D]. 南京中医药大学, 2015.

知识发现与应用

本章的写作及润色得到了南京大学计算机科学与技术系及智能信息处理研究组的吴骏博士及彭岳、杨骏、戴恒宇、李红、王茜、陈厚兵、陆恒杨、谢路遥等几位同学的协助，在此表示深深的谢意。

9.1 引言

起初，神创造天地；第一日造光；第二日造空气；第三日造蔬菜及结果子的树木；第四日造白天和夜晚；第五日造一切带有生命各从其类的动物；第六日按照神的样式造人……（摘自《圣经·创世纪》）

我们无从知晓这些先知如何知道“神创世纪”，我们也无须知晓这个神是谁造的，总之冥冥之中，人类起初都是在接受这些具有宗教、哲学意义的认知基础上，观察人类所生存的世界，并通过各种演绎方法获悉对这个世界的更多理解。在整个认知的过程中，根据实践及来自实践的反馈归纳出越来越多的“先天法则”，并利用这些先天法则继续演绎，如此反复。

上述过程其实就是人类认识自然、改造自然的过程。在这个过程中，人类遇到了来自方方面面的挑战和困难，出于人类生存的最原始需求，人类不断地追求思维认识层次、技术手段层次、工具开发层次等的提高，或许也正因为如此，人类本身才能得到不断的进步。而现在，我们正处在一个被标签化为“大数据”的时代，对于我们而言：这是新的机遇，也是新的挑战。

从应用的角度来看，大数据的价值在于有用，这也是“政产学研商用”各界对其抱有极大期望的根本原因，这也正是大数据的“4V”特征中所表示的“Value”属性。但是特别要注意的是，持有这种观点一定是基于独立的第三方的全局视角，而针对一个具体的应用，鉴于大数据在数据规模上的巨大，其价值密度往往是稀疏的，甚至我们没有任何判定方法去客观评判收集到的数据对目标应用是否直接有用。

有一个听起来很奇怪的故事可以说明大数据应用在数据获取时所面临的问题，故事原型是这样的：有一个醉汉在昏暗的路灯下找东西，一位路人经过，问他：你在找什么？醉汉答：在找我的车钥匙，不小心弄丢了。于是，路人又问：你是在这边弄丢的吗？醉汉答：在暗处丢的。路人不解地问：那为什么在这里找呢？醉汉说：因为这里相对比较亮。

看起来啼笑皆非的故事事实上在大数据分析的场景下也会经常遇到：当我们尝试去做一个目标驱动的应用时，基于数据驱动的思维，首先要做的就是数据的收集，而我们往往会认为（其实是一种误解或无奈）：利用已收集到的及通过可行的手段努力收集到的数据就能够进行目标求解。不过需要理性考虑的是：能够收集到的这些数据对目标问题的解决是否有用呢？进一步需要考虑的是，为了确保收集到的数据能够满足目标问题求解的需求，我们会收集巨量的数据，而在巨量数据中对目标有用的那部分数据（往往极其稀疏），是否真的有机会被甄选出来？前者是大数据项目建设的论域问题，后者涉及大数据项目的技术选型问题。

大数据在应用层的挑战主要包括数据规模极其巨大、数据类型极其复杂、数据更新极其快速，这些挑战分别体现在数据存储、数据计算、系统运维等不同环节。从技术选型的角度来看，从各种候选的技术手段中选出最适合的一个（种）应是最理性的做法，而其中隐含的问题是：大数据场景下的应用问题是一个开环问题，这就意味着根据历史数据情况选择的精度和性能均表现优异的算法，在新增的数据场景下未必依然表现优异，甚至未必可接受，这就意味着算法选型会因为问题场景的变化而不断迭代变化（重新选择另外的技术手段，或者在现有的技术手段上进行技术改进，前者其实就是面对需求膨胀做改变的话题，是工程中最为忌讳的，后者对应于技术预研，且不谈需要投入的人力成本，时间成本往往也是难以控制的）；另一方面，每一个技术选型（比如数据库选型）往往意味着经费的投入及配套人力的跟进（包括开发和运维），而一个项目的建设总是有成本约束的，这在工程管理上也是一个难题。

从技术的角度来看，大数据的核心流程是：“数据采集层→数据存取层→数据分析层→系统运维层”，大数据的本质属性导致上述流程的各个环节都充满了挑战和困难，比如：

（1）数据采集层的挑战

大数据精神是要求尽可能地采集更多数据源的数据，这隐含着几个问题（包括但不限于）：①研发更多的数据采集设备进行更多的数据采集，比如在前面第3章大数据产业中提到的个人健康数据的采集，就是尝试开发更多的设备，从多种渠道收集人的健康数据。②从更多的数据源采集更多的数据，而数据的价值在于数据之间能够建立“连接”，这就意味着需要对各种数据源的数据加以整合，往往会牵涉到数据本身及论域分析的话题，这是个不可回避的难题。③数据格式的多样性导致如何有效地表示和理解这些异构的数据成为一个技术难题，比如文本理解、图像理解、视频理解等（当然这将涉及后续的数据分析）。④多源数据的采集和整合带来的一个直接挑战是如何对这些数据进行预处理及进行有效的质量监管，并且这样的质量监管要持续存在于数据使用的整个生命周期。

（2）数据存取层的挑战

数据存取层的本质目标是数据独立性、格式规范性、操作统一性、平台透明性、运行高

效性,大数据场景下,数据规模的海量性,以及由此引发的数据海量吞吐率需求对数据的组织与管理方式带来了极大的挑战,前文提及的各种类型的数据库,包括 SQL、NoSQL、NewSQL,本质上都是在数据存取层面响应类似的挑战。

(3) 数据分析层的挑战

大数据复杂的数据关联性以及数据规模的海量性导致在小数据集上有效的传统算法失效、计算复杂度过高(以至于现有的计算无法承受)等问题,鉴于数据分析是大数据的核心,因此这类挑战几乎无法回避,现有的一些响应策略包括:通过并行化分而治之或并行化处理以解决复杂问题的求解,或者寻找新算法、新理论以降低复杂问题的计算复杂度、寻求与数据尺度无关的近似算法以提高计算的可行性等。

鉴于机器学习的天然使命(机器学习生来就是通过数据自动学习得到知识,见 9.2.3 节),以及大数据分析的天然适应性,机器学习已成为大数据分析中不可或缺的关键技术,而传统的机器学习一般是面向小数据集的(面临着单点式挖掘、依存度高、实时性差等困难,难以应对大规模数据及分布式存储,也难以对快速变化的数据集进行实时处理),因此,以机器学习为中心进行策略和算法的改进,使其适合于大数据应用场景,是迎接大数据挑战的重要举措,主流的策略和手段包括:分布式机器学习、并行机器学习、基于 GPU(加速)的机器学习和深度学习等。

理想情况下,数据越大,机器学习可以训练的模型越复杂,模型的能力就越强大,然而这是针对小数据场景的。大数据场景下,大规模的数据给机器学习造成了计算上的困难,尤其是对于复杂的模型,计算量可能会随着数据量超线性增长。因此在很多大数据应用场景下,往往是使用简单的模型(比如回归)而不是使用更复杂的模型(比如 SVM)来对规模巨大的数据进行处理,这其实是不得已而为之。用大数据训练复杂模型,最主要的困难就是内存容量限制和计算时间过长,为了解决这两个问题,必须从算法层面上突破计算量和内存存储量的限制。通常有三种策略:并行算法、在线算法和近似算法。

所谓并行算法,就是采用“分而治之”的方式,将计算任务拆解成多个小的任务,并将小任务分配到多个处理器上进行计算从而使得机器学习的算法更快,而这种处理策略对硬件条件有较高的要求,需要 MapReduce 集群、MPI 集群、GPU 集群等计算资源。

传统的机器学习算法通常每一轮都用全部的数据来更新模型,而在线算法则有所不同,每次只获取少量数据来更新模型。这样做的好处是内存消耗大大减小,而且每一轮的计算量也很小。在线算法(或随机算法)通常比较适用于单机的计算环境,对计算资源要求较小,即使没有强大的集群也可以用大数据来训练机器学习模型。

很多机器学习问题可以归结为矩阵分解的问题,而这些运算的时间和内存消耗通常正比于数据量的平方。如果不能高效求解矩阵运算,很多机器学习方法就无法应用于大数据问题。随机近似是目前处理大规模矩阵分解的最有效方法,即快速求出一个子矩阵,使得子矩阵尽量多地保留原矩阵的某种性质,然后对子矩阵进行矩阵分解等操作,最后用子矩阵的分解结果近似地作为大矩阵的分解结果,现在随机投影、Nystrom 方法等随机近似算法已经成为处理

大规模矩阵分解最有效的方法。

为了提升特定数值运算操作（如矩阵相乘、矩阵相加、矩阵—向量乘法等）的速度，向量化编程是提高算法速度的一种有效方法。向量化编程强调单一指令并行操作多条相似数据，形成单指令多数据流（SIMD）的编程泛型。然而，在单个 CPU 上执行时，向量运算会被展开成循环的形式，本质上还是串行执行。GPU (Graphic Process Unit, 图形处理器) 的众核体系结构包含几千个流处理器，可将向量运算并行化执行，大幅缩短计算时间。随着 NVIDIA、AMD 等公司不断推进其 GPU 的大规模并行架构支持，面向通用计算的 GPU (General-Purposed GPU, GPGPU) 已成为加速可并行应用程序的重要手段。

深度学习被认为是最接近人脑的智能学习方法，近些年得到学术界和工业界的普遍追捧。深度学习采用的模型为深层（非线性）神经网络（Deep Neural Network, DNN）模型，通过使用包含多个隐藏层（Hidden Layer, 也称隐含层）的神经网络（Neural Network, NN）实现复杂函数逼近，并展现了强大的学习数据集本质和高度抽象化特征的能力。与传统的浅层模型相比，深层模型经过了若干层非线性变换，带给模型强大的表达能力，从而有条件为更复杂的任务建模。与人工特征工程相比，深度学习不仅能够自动学习特征，而且还能挖掘出数据中丰富的内在信息，并具备更强的可扩展性，顺应了大数据的趋势。有了充足的训练样本，复杂的深层模型可以充分发挥其潜力，挖掘出海量数据中蕴含的丰富信息。而基于强有力的基础设施和定制化的并行计算框架，让以往不可想象的训练任务加速完成，从而为深度学习走向实用奠定了坚实的基础。

（4）系统运维（架构）层的挑战

系统运维（架构）层的两个主要任务是：①提供各种计算流（数据采集、数据存储、数据分析等）的高效流转和运行。②提供面向用户的统一服务平台。前者专注于计算架构的技术选型，后者专注于系统的部署模式。如前所述，为了响应大数据大吞吐率的需求，计算性能的提高是“众望所归”，分布式架构及并行架构几乎成为必然，目前可选型的计算架构包括 Hadoop、Spark、Storm 等，在大数据应用场景下，根据应用场景的特点（数据特点及应用特点）进行计算架构选择之后，需要关注两个问题，即上述的所有计算流涉及的各种计算和算法需要进行因为采纳这种而非那种计算架构（模型）的算法改进；同时要确保计算架构本身的稳定性和适应性，这又对应于计算架构和模型的改进和完善。就部署模式方面而言，云模式几乎成为大数据运营的事实标准，此处不再赘述。

从科学的层面来看，大数据背后隐含着若干关键的科学问题，比如可计算问题。计算机科学关注的是可计算问题，而传统的可计算问题可以归结为算法问题，即如果此算法不能图灵可计算，即可判定该问题（计算）无法由计算机处理。而传统的计算机科学关于计算的研究是专注于计算本身，而不在意数据，或者说传统的计算是基于“数据完备、不变”的假设，而大数据场景下，数据是持续变化和更新的，这也意味着新的问题：传统的计算理论在大数据场景下是否还有效？或者说，在大数据场景下，针对数据海量且不断增量的事实，这个问题是否可计算？进而，在可计算问题之后还可延伸的若干问题，比如计算可信问题、资源

(数据、计算、能源等)管理问题等,所有这些典型的计算机科学问题,在大数据场景下,大家更愿意用数据科学来进行描述。

综上所述,针对大数据带来的机遇和若干挑战,不同角色的工作者从不同角度、利用不同的思维和策略给予了积极的响应和拥抱:大数据应用开发者以价值驱动的方式开发面向目标需求的落地应用,并以成功部署实施来获得期望价值的方法来体现出大数据应用者的价值;大数据技术研究者以问题驱动的方式研究面向大数据细分领域(需求)的关键技术,并因此技术成果的普适性、泛化性和稳定性获得价值认可;大数据科学研究者以数据为研究对象尝试揭示数据的可计算、可管理及计算可信问题,数据科学的研究本身并不关心数据的价值,但因为此科学对所有围绕大数据展开的相关技术研究及工程开发具有普适的指导意义,而彰显出大数据科学者(群体)的价值;或许还有哲学研究者,会从哲学的层面去研究诸如“大数据的本质是什么”、“大数据从哪里来”、“大数据到哪里去”等诸多命题。

本章尝试从哲学、科学、技术、工程等视角梳理拥有不同知识背景和价值期望的利益主体响应和拥抱大数据的思维方式及若干具体行动,本章下面的结构安排如下:9.2节简单地介绍与数据分析紧密相关的若干技术,尝试在梳理人工智能研究的历史脉络基础上,阐述机器学习的位置和角色,并着重介绍机器学习在数据分析领域的应用,即数据挖掘的目标、分类及一般技术路径;9.3节在简述哲学、科学、技术及工程应用的基础上,介绍大数据给研究者带来的两个变革,包括新型科学发现研究范式的提出,以及作为一门学科体系提出的数据科学;9.4节从方法论的角度分析数据挖掘在实际应用中,大数据工作者应该具备并时刻考虑的思维方式;9.5节对本章进行小结。

9.2 从机器学习到数据挖掘

在大数据俨然成为一个“政产学研商用”各界普遍关注的热点,甚至媒体也在以不同方式热炒的当代,“数据挖掘”作为一个技术名词越来越得到各界的重视和认可,甚至是盲目的膜拜。这或许是因为“大数据”和“数据挖掘”都兼有“数据”这个名词,也或者是因为人们对大数据的价值期望最终希望通过一门落地的技术加以实现,而“数据挖掘”恰是人们认可的、而事实上也确实会在大数据应用中有所担当的一种手段。从技术层次上来看,数据挖掘就是一种从数据中寻找规律的技术。

数据挖掘受到了很多学科领域的影响,其中数据库、机器学习、统计学的影响无疑最大,而在谈论数据挖掘和各个学科之间关系的时候,“人工智能”作为一门学科不得不被提及,因为,几乎得到普遍认同的一个观点是:数据挖掘是人工智能的一个方向(或者分支)。至于在未来,数据挖掘是否会由于突飞猛进的发展而独立成为一个新的学科,那就是另外一回事了。

在前文第6章“数据存储与管理”中,专门就数据库进行过讨论和分析,而且,在进行数据挖掘的实际操作中,往往会默认已经有成熟的数据库平台作为支撑,至于技术选型中因为采纳了这种或那种的数据库而使得数据挖掘算法必须进行适应性改进,那是实现层次的事

(当然,这也很重要),本节不予关心。本节将重点讨论和分析机器学习、数据挖掘、统计学及人工智能这4个技术名词(学科)。

9.2.1 统计与统计学

统计学是一门研究怎样收集、组织、分析和解释数据中数字化信息的科学,统计学可以分为两大类:描述统计学和推断统计学,前者涉及组织、累加和描绘数据中的信息,后者涉及使用抽样数据来推断总体。其应用几乎覆盖了社会科学和自然科学的各个领域。

统计学的最早研究可追溯至古希腊的亚里士多德时代,在两千多年的发展过程中,统计学至少经历了“城邦政情”“政治算术”“数理统计学”“社会统计学”“统计分析科学”等几个发展阶段(学术流派):

1) 城邦政情(Matters of state)阶段始于古希腊的亚里士多德撰写的“城邦政情”或“城邦纪要”(类似于社会科学研究的输出文档),这些“城邦政情”或“城邦纪要”记录的内容主要包括各城邦的历史、行政、科学、艺术、人口、资源和财富等社会和经济情况的比较和分析。

2) 1690年,英国威廉·配第(William Petty,被誉为“近代统计学之父”)撰写和出版的《政治算术》是“政治算术”阶段开始的标志,“政治算术”的特点是将统计方法与数学计算和推理方法相结合,更注重用定量分析的方法分析社会经济问题。在这部书中,威廉·配第运用统计方法对英国、法国和荷兰三国的国情国力,进行了系统的数量对比分析,从而为统计学的形成和发展奠定了方法论基础。

3) 18世纪末至19世纪末是统计学的发展时期,这个时期形成了两大主要学派,分别是数理统计学派和社会统计学派,前者将概率论引进统计学而形成数理学派,代表性人物是比利时人阿道夫·凯特勒(Adolphe Quetelet),其主张用研究自然科学的方法研究社会现象,正式把古典概率论引进统计学,使统计学在“政治算术”所建立的“算术”基础上,又在准确化的道路上大大地跨进了一步,为数理统计学的形成与发展奠定了基础,被誉为国际统计会议之父、近代统计学之父、数理统计学派创始人。

4) 社会统计学派产生于19世纪后半叶,创始人是德国统计学家卡尔·古斯塔夫·阿道夫·克里斯(Karl Gustav Adolf Knies),他提出统计学是一门独立的社会科学,是一门对社会经济现象进行数量对比分析的科学。代表性人物还有德国统计学家和经济学家恩斯特·恩格尔(Ernst Engel,其提出的“恩格尔系数”至今仍被广泛使用)、美国经济学家西蒙·史密斯·库兹涅茨(Simon Smith Kuznets,提出库兹涅茨周期理论及国民收入核算理论,被誉为“美国的G. N. P.之父”)、英国经济学家约翰·理查德·尼古拉斯·斯通(John Richard Nicholas Stone,修订了联合国《国民经济核算体系及辅助表》,成功推出了联合国《国民经济核算体系(1968年)》,并为联合国制定了《社会和人口体系》等)。

5) 19世纪末,欧洲大学开设的“国情纪要”或“政治算术”等课程名称逐渐消失,代之而起的是“统计分析科学”(science of statistical analysis)课程,这一课程的出现是现代统

计学发展阶段的开端。威廉·斯利·戈塞特 (William Slessy Gosset) 在 1908 年发表的关于 t 分布的论文是统计学发展史上划时代的文章, 创立了小样本代替大样本的方法, 开创了统计学的新纪元。

统计学的英文单词 “statistics”, 最早是由戈特弗里德·阿亨瓦尔 (Gottfried Achenwall) 于 1749 年开始使用, 其词根起源或许是亚里士多德时代的 “Matters of state” (城邦政情)、现代拉丁文的 “Statisticum collegium” (国会) 及意大利文 “Statista” (国民或政治家)。统计学是研究如何测定、收集、整理、归纳和分析反映客观现象总体数量的数据, 以便给出正确认识的方法论科学, 被广泛地应用在各门学科之上, 早期用来表征国家的经济水平及用于军事用途的物质资源, 随后统计学的用途扩展到数据的分析及其组织, 其应用领域从自然科学和社会科学到人文科学, 甚至被用于工商业及政府的情报决策之上。

目前来看, 统计学是应用数学的一个分支, 主要利用概率论建立数学模型, 收集所观察系统的数据, 进行量化分析、总结, 做出推断和预测, 为相关决策提供依据和参考。统计学之于本文所关心的数据驱动相关的问题, 其贡献至少在于 (包括但不限于):

1) 演绎逻辑和归纳逻辑是人类的两种基本思考方式, 而归纳逻辑的思维方式是机器学习与数据挖掘的基本方法论。在归纳逻辑的研究历程中, 古典归纳逻辑向现代归纳逻辑演变的一个标志是: 是否归纳前提与结论间的概然性。代表性人物凯恩斯 (John Maynard Keynes) 将统计概率看作两命题或命题集合之间的一种逻辑关系, 从而建立了第一个公理化的概率演算系统并创建了第一个现代归纳逻辑的系统理论, 这对于理论计算机、人工智能、机器学习、数据挖掘等相关学科和技术的发展具有极大的理论意义和现实价值。

2) 作为一门研究如何测定、收集、整理、归纳和分析数据的方法论科学, 统计学从数据采集到最终分析应用积累了大量的理论基础、数学模型、计算工具等, 已经能够有效地进行分析, 而从统计学的发展历史也能看出统计学在数据分析方面的实践意义和价值。

3) 统计学本身在数据分析领域的功用得到了广泛的认可, 有很多厂商专门开发了面向统计分析的软件产品, 主流产品如 SAS (Statistical Analysis System)、SPSS (Statistical Package for the Social Science)、Excel (微软公司推出的 Office 系列产品之一, 本身具有简单的统计功能, 可配载功能强大的数据分析插件 XLSTAT)、R (一种用于统计分析、绘图的开源语言和操作环境, 大数据时代得到很大的关注热度)、Stata 统计软件 (美国计算机资源中心研制)、Minitab (美国宾州大学研制) 等, 这些软件产品的广泛使用进一步丰富和扩展了数据分析应用的技术选型和产品选型范围。

4) 统计学为机器学习的发展夯实了数学理论基础, 机器学习中的很多理论基础来源于统计, 甚至有统计学家认为: 机器学习是统计学的一个研究 (应用) 分支。

9.2.2 智能与人工智能

从感觉到记忆再到思维这一过程, 称为 “智慧”, 智慧的结果就是产生了行为和语言, 将行为和语言的表达过程称为 “能力”, 两者合称 “智能”。智能的英译是 “Intelligence”, 其词

根来源是拉丁文的“Legere”和“Intelegere”，前者指采集（特别是果实）、收集、汇集，并由此进行选择，形成一个东西；后者是指从中选择，进行理解、领悟和认识。

人工智能（Artificial Intelligence, AI），也称人造智能，是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。不过类似的定义并没有一个官方的统一认可，有很多关于人工智能的定义和描述，比如：与人的思维相关的活动，诸如决策、问题求解和学习等的自动化；一种使计算机能够进行思维活动，使机器具有智力的激动人心的新尝试；研究如何让计算机做现阶段只有人才能做得好的事情；那些使知觉、推理和行为成为可能的计算的研究；关于人造物的智能行为，包括知觉、推理、学习、交流和在复杂环境中的行为；像人一样（理性）思考、（理性）行动的系统等。

1956年之前，后来被人工智能吸收的各种理论、思维方式和技术都得到了充分的发展，比如古希腊哲学家和思想家亚里士多德创立的至今仍在沿用的演绎法（三段论）、德国数学家和哲学家莱布尼茨（Gottfried Wilhelm Leibniz）创立的为后来数理逻辑的发展奠定基础的形式逻辑符号化、被誉为“人工智能之父”的英国数学家图灵（Alan Mathison Turing）在1950年提出的“机器也能思维”的论断、美国神经学家迈科洛奇和皮兹在1943年建成的第一个神经网络模型（MP模型）、美国数学家维纳（Norbert Wiener，控制论创始人）在1948年创立的对后来人工智能影响巨大并形成了行为主义学派的控制论等。

1956年夏季，约翰·麦卡锡（John McCarthy）、明斯基（Marvin Minsky）、罗切斯特（Nathaniel Rochester）和香农（Claude Shannon）等共同发起，并邀请莫尔（Trenchard More）、塞缪尔（Arthur Samuel）、纽厄尔（Allen Newell）和西蒙（Herbert A. Simon）等学者参加达特茅斯会议（Dartmouth Conference），共同研究和探讨用机器模拟智能的一系列有关问题，并首次提出了“人工智能”，它标志着“人工智能”这门新兴学科的正式诞生。

1956年以后，研究者们发展了众多理论和原理，人工智能的概念也随之扩展，并且它们也在影响着其他技术的发展。

在人工智能的定义和理解方面，一直有两个学派：一个是强人工智能（Bottom-up AI），该学派认为有可能制造出能真正推理（Reasoning）和解决问题（Problem Solving）的智能机器，并且这样的机器智能被认为是具有知觉的和有自我意识的；另一个是弱人工智能（Top-Down AI），该学派认为不可能制造出能真正推理和解决问题的智能机器，这些机器只不过看起来像是智能的，但是并不真正拥有智能，也不会有自主意识。约翰·麦卡锡在1956年达特茅斯会议上提出的人工智能的定义（“人工智能是要让机器的行为看起来就像是人所表现出的智能行为一样”）事实上是弱人工智能学派的定义。主流科研集中在弱人工智能上，并且一般认为这一研究领域已经取得了可观的成就，强人工智能的研究则处于停滞不前的状态下。

人工智能是一门综合性的交叉学科和边缘学科，涉及的学科领域包括（但不限于）计算机科学与技术、信息处理和自动化技术、智能科学、认知科学、心理科学、脑及神经科学、

生命科学、语言学、逻辑学、行为科学、教育科学、系统科学、数理科学及控制论、哲学甚至经济学等众多学科领域。所有关于人工智能的研究可以归结为以下4点:

1) 1个(研究)目标: 人工智能的研究目标就是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统, 由于目前实现的载体都是计算机, 因此人工智能与计算机软件有密切的关系。

2) 2个(研究)角度: 机器智能及智能机器是人工智能的不同研究视角, 前者关注如何让机器具有(人一样的)智能, 更接近于弱人工智能的定义; 后者关注如何实现具有(人一样的)智能的机器, 更接近于强人工智能的定义。

3) 3个(研究)层次: 思维、感知、行为是人工智能研究者的3个主要研究层次, 其主要目标是实现“像人一样感知环境”、“像人一样思考决策”、“像人一样行动”, 值得注意的是: 这里的“行动”应广义地理解为采取行动或制定行动的决策, 而不仅是肢体动作。

4) 4种(科学)模拟: 结构模拟、功能模拟、行为模拟、机制模拟是人工智能研究者的4种主要的研究(检验)思路。

人工智能的研究内容(分支)涉及面很广, 至少包括(但不限于): 探索与求解、学习与发现、知识与推理、发明与创造、感知与交流、记忆与联想、系统与建造、应用与工程等。机器学习这一研究领域是AI的一个分支, 用来帮助机器和软件进行自我学习以解决遇到的问题, 对其的研究可追溯至1950年被誉为“人工智能之父”的英国数学家图灵提出的“机器也能思维”的论断。

9.2.3 人工智能与机器学习

机器学习作为一个独立的分支得到重视和发展, 事实上是与人工智能这一学科的发展紧密耦合在一起的, 并且因为人们对智能的认知改变(演化), 而逐步被提升到白热化的关注度水平。

“人工智能”这个概念从1956年正式推出即得到研究者的普遍关注, 从20世纪50年代到70年代初, 人们认为只要给机器赋予逻辑推理的能力, 机器就能具有智能, 因此这个阶段的有关人工智能的研究被标签化为“推理期人工智能”。

随着研究的不断推进, 人们逐渐认识到, 仅具有逻辑推理能力是远远实现不了人工智能的。基于“要使机器具有智能, 就必须设法使机器拥有知识”的价值理念, 在费根鲍姆(E. A. Feigenbaum, 被誉为“知识工程之父”, 获得1994年的图灵奖)的倡导下, 20世纪70年代中期开始, 人工智能进入了被标签化为“知识期人工智能”的阶段, 在这一时期, 大量专家系统问世, 在很多领域都做出了巨大的贡献。

随着研究的进一步推进, 专家系统面临“知识工程瓶颈”, 研究者发现由人将知识总结出来再教给计算机是相当困难的, 因此一个自然的想法是: 能否让机器自己学习知识? 事实上, 机器学习在20世纪80年代正式被视为“解决知识工程瓶颈问题的关键”而走到人工智能主舞台的聚光灯下的。应该注意到的是: 1950年图灵在图灵测试的文章中, 已经提到了机器学

习的可能,而关于机器学习的研究工作,从20世纪50年代起已经展开,主要集中在基于神经网络的连接主义学习方面,在20世纪60年代到70年代,多种学习技术得到了初步发展,例如以决策理论为基础的统计学习技术及强化学习技术等,代表性工作主要有塞缪尔(A. L. Samuel)的跳棋程序及尼尔森(N. J. Nilson)的“学习机器”等。可以说,统计学习理论的许多重要结果也是在这个时期取得的。总的来说,20世纪80年代是机器学习成为一个独立的学科领域并开始快速发展、各种机器学习技术百花齐放的时期。1983年,由米哈尔斯基(Ryszard S. Michalski)、卡博奈尔(Jaime G. Carbonell)和汤姆·米切尔(Tom Mitchell)合作出版的《Machine Learning: An Artificial Intelligence Approach》一书标志着机器学习正式成为人工智能中一个独立的领域(该书其实是一部集早期机器学习研究之大成的文集)。

机器学习的研究者出于不同的价值观和研究视角,衍生出了很多的学术派别,其中最为著名的有两个(而这两个流派的价值观点几乎是针锋相对的):

1) 把机器学习看作“人工智能”的一个分支,抱有这样价值观点的主体往往是计算机科学家。人工智能出身的大多数机器学习研究者往往是把机器学习作为实现人工智能的一个途径。他们专注于以机器学习为手段解决人工智能中的问题,但具体采用什么样的学习手段(基于统计的、代数的、逻辑的、几何的),他们并不关心。他们关心的是:人工智能问题或许并不是数学问题,甚至未必是依靠数学才能够解决的问题,人工智能中许多事情的难处,往往在于我们并不知道困难的本质在哪里,不知道“问题”在哪里。一旦“问题”清楚了,解决起来可能并不困难。

2) 把机器学习看作“应用统计学”的一个分支,抱有这样价值观点的主体往往是统计学家。统计学出身的机器学习研究者,绝大部分是把机器学习当作应用统计学。他们专注于如何把统计学中的理论和方法变成可以在计算机上有效实现的算法,至于这样的算法对人工智能中的什么问题有用,他们并不关心。这群人或许对人工智能毫无兴趣,在他们眼中,机器学习就是统计学习,是统计学比较偏向应用的一个分支,充其量是统计学与计算机科学的交叉。由于统计学习之外的学习手段(比如基于代数的、逻辑的、几何的)难以纳入统计学的范畴,这个群体往往是被传统统计学家排斥的。

根据不同的分类标准,机器学习可以分为不同的类别,比如:按照学习策略(指学习过程中系统所采用的推理策略)进行分类,机器学习可分为:机械学习、示例学习、演绎学习、类比学习、基于解释的学习、归纳学习等;按照所获取知识的表示形式进行分类,可分为:代数表达式参数、决策树、形式文法、产生式规则、形式逻辑表达式、图和网络、框架和模式、神经网络等;基于应用领域进行分类,机器学习最主要的应用领域有:专家系统、认知模拟、规划和问题求解、数据挖掘、网络信息服务、图像识别、故障诊断、自然语言理解、机器人和博弈等;综合考虑各种学习方法出现的历史渊源、知识表示、推理策略、结果评估的相似性、研究人员交流的相对集中性及应用领域等诸条因素,可将机器学习方法分为:经验性归纳学习、分析学习、类比学习、遗传算法、连接学习、强化学习等;按照学习的形式

进行分类,机器学习可分为监督学习、非监督学习和半监督学习等。

连接主义学习技术在20世纪50年代曾经经历了一个大发展时期,但因为早期的很多人工智能研究者对符号表示有特别的偏爱,因此当时连接主义的研究并没有被纳入主流人工智能的范畴。同时,连接主义学习自身也遇到了极大的问题,明斯基(M. Minsky)和西蒙·派珀特(S. Papert)在1969年指出,(当时的)神经网络只能用于线性分类,对哪怕“异或”这么简单的问题都解决不了。于是,连接主义学习在此后近15年的时间内陷入了停滞期。直到1983年,霍普菲尔德(J. J. Hopfield)利用神经网络求解TSP的问题获得了成功,才使得连接主义重新受到人们的关注。1986年,鲁梅尔哈特(D. E. Rumelhart)、辛顿(G. E. Hinton)和威廉姆斯(R. J. Williams)发明了著名的BP神经网络算法(按误差逆传播算法训练的多层前馈网络),该算法可以说是最成功的神经网络学习算法,在当时迅速成为最流行的算法,并在很多应用中都取得了极大的成功。

基于连接主义的学习常被人诟病的一个问题是其“试错性”,即:在此类技术中有大量的经验参数需要设置,例如神经网络的隐层结点数、学习率等,在实际工程应用中,人们可以通过调试来确定较好的参数设置,但对机器学习研究者来说,对此显然是不满意的。或许正是由于这个原因,研究者开始把目光转向了统计学习,并从20世纪90年代开始,统计学习逐渐成为机器学习的主流技术。

其实早在20世纪60年代到70年代就已经有统计学习方面的研究工作,统计学习理论在那个时期已经打下了基础,如弗拉基米尔·万普尼克(V. N. Vapnik)提出的“支持向量机”概念(1963年)、VC维(1968年,合作者是亚历克塞·泽范兰杰斯, A. J. Chervonenkis)、结构风险最小化原则(1974年)等。在此基础上,伯恩哈德·宝狮(B. E. Boser)、盖伊恩(I. Guyon)和弗拉基米尔·万普尼克在1995年提出有效的支持向量机算法,尤阿基姆(T. Joachims)等人在文本分类的研究中凸显优势,统计学习的优势逐渐显现。然而,统计学习也有其局限,比如:①理论上可以通过把原始空间利用核技巧转化到一个新的特征空间使得困难的问题得到简化,但在实际操作中选择合适的核映射需要浓厚的经验色彩。②统计学习与连接主义学习技术都是基于“属性-值”的表示形式,难以有效地表示出复杂数据和复杂的数据关系,不仅难以利用领域知识,而且学习结果还具有“黑箱性”。③传统的统计学习技术往往因为要确保统计性质或简化问题而做出一些假设,然而在实际操作中,这些假设往往是不现实的。如何克服这些缺陷,正是很多学者正在关注的问题。

9.2.4 数据挖掘及技术路径

数据挖掘(Data Mining, DM)一般是指从大量的数据中通过算法搜索隐藏于其中的信息的过程,在很多场合往往与数据库知识发现(Knowledge Discovery in Databases, KDD)通用,但更为一般的理解是:数据挖掘是知识发现过程的一个关键步骤。

数据库知识发现,更一般的称谓就是知识发现,所谓知识发现指的是用一种简洁的方式

从大量数据中抽取信息的一种技术，所抽取的信息是隐含的、未知的，并且具有潜在的应用价值。根据这样的定义和理解，可以看出：知识发现的操作对象是数据（库），目标是发现有价值的信息，这些信息表示不同研究对象之间的关系和模式（知识）。因此有关知识发现的研究一般涉及知识表示、知识发现方法、知识发现应用等。

知识发现的过程一般包括三个主要阶段，分别是：数据准备、数据挖掘、解释和评估。

1) 数据准备阶段主要从数据库中提取出相关数据，并对数据进行集成、选择和预处理，从而（更好地）满足后续数据挖掘阶段的数据输出。

2) 数据挖掘阶段主要利用相关方法对上一阶段的输入数据进行特征提取与选择、数据建模，从而达到从数据中发现关系和模式的目的。

3) 解释和评估阶段主要对数据挖掘阶段的输出进行客观分析和评估，这一步骤往往会耦合到第二阶段，这是因为：①如果建模结果不满足用户的要求，需要更换新的挖掘方法进行再尝试。②如果建模的结果过于抽象，用户不容易理解，需要更换新的挖掘方法进行再尝试。③如果所有采用的数据挖掘技术在精准度等性能指标上不满足用户的需求，也需要更换新的挖掘方法进行再尝试。

从知识发现的一般流程描述可以看到，很多人愿意将数据挖掘和知识发现广义上理解为一件事情，本质的原因就在于：以数据挖掘为核心的知识发现，其关键步骤几乎是完全相同的。

数据挖掘的基础是数据，所谓数据指的是数据对象及其属性的集合，属性指的是实体对象的特征（向量），比如人的年龄、肤色、性别等，而所有的这些特征（向量）描述的就是一个（数据）对象，根据数据对象和数据类型的不同，数据挖掘可以分为以下几种（包括但不限于）：

(1) 面向“属性-值”的数据挖掘

大多数情况下，一个数据对象的数据可以格式化为“属性-值”的模式，特别是针对关系型数据库的数据（见表9-1），每一个记录描述的是具体某一个人（比如王小明）各门功课的成绩及（老师给的）综合评定。在如表9-1所示的结构化数据中，每一条记录（可以视为一个数据对象）的每一个字段（可以视为一个属性）就是一个特征，所有特征组合在一起就是一个特征向量，而字段“综合评定”代表的就是“值”。

表9-1 学生成绩表示意

ID	姓名	语文	数学	英语	综合评定
1	王小明	90	85	85		优
2	谢小明	88	80	85		良
3	张小明	75	80	85		中
4	吴小明	90	85	85		中
5	李小明	90	85	85		优
6	陈小明	80	85	85		良
.....					

特别值得一提的是：“属性-值”数据形式中的“值”不是必需的，比如：数据挖掘中的聚类算法，仅就属性特征展开，根据属性特征（向量）的相似性进行聚集；数据挖掘中的关联规则挖掘算法，也仅就属性特征展开，根据属性与属性之间的“常联系关系”进行分析。当然在关联规则挖掘中，也可以利用“值”的信息，比如根据不同“值”意义下的数据集进行挖掘，不过这已经是处理策略的问题了。

数据挖掘中的分类算法，其对数据的要求是必须要有“值”这个数据的，这个“值”代表着领域专家对具有对应的“属性特征”的一个判定，这个判定可以是连续值，也可以是离散值，前者对应于回归问题，后者对应于分类问题。

还需要说明的一点是：在很多参考书或应用场景下，研究者往往会将此处的“值”用“类标属性”或“值属性”来表示，前者对应于分类问题中的“类标”，后者对应于回归问题中的“目标值”，而与这两个称呼对应的“属性”一般称为“特征属性”。

（2）面向“链接型”数据对象的数据挖掘

不是所有的数据对象都可以简单地格式化为上述“属性-值”的形式，比如描述人与人社会关系的社会网络、Web 网页（每个网页都会有 URL 扇出，或者被另外的 URL 扇入）等，这些数据对象本身描述的是一个以“节点”和“关系”为主体形成的图（网络），因此，针对这些图的挖掘必须有专门的数据挖掘手段加以应对，这就是社会网络分析或链接型数据挖掘。根据社会网络的“链接性”特点，社会网络分析的研究对象又包括：针对节点的研究、针对关系的研究、针对社团的研究、针对网络拓扑（社团）演化的研究、针对信息传播的研究等（关于社会网络分析的内容，本文不再赘述）。值得注意的是，链接型数据挖掘，或者社会网络分析，与普通的“属性-值”数据挖掘并不是完全独立的，比如：①针对节点（或者关系）的研究，就可以把每个节点（关系）建模成一个“属性-值”的数据对象。②针对“属性-值”的数据挖掘算法中，有一类是做关联规则发现的，其事实上就是将每个属性看作一个数据对象，而建立起以“属性”为节点，属性与属性的“共同出现”为“关系”的图。或者更为理性地说：一个数据对象是建模为“属性-值”型还是建模为“链接型”，本质上仅仅是建模的出发点不同而已。

数据挖掘在社会网络分析中的应用主要包括（但不限于）基于链接的节点排序、基于链接的节点分类、节点聚类、链接预测、子图发现和图分类，具体而言：①基于链接的节点排序指的是利用图中的链接结构，根据某种衡量节点重要性的度量来对图中的节点进行排序。②基于链接的节点分类与传统分类问题最显著的不同在于节点的类别是彼此相关的，如何设计合理的分类算法才能有效地利用这些相关信息是研究者所面临的挑战。③节点聚类又称为群体检测，它的目的是将有着共同特征的节点聚类。④链接预测是基于所链接的节点属性和已观测到的链接来预测某链接是否存在。⑤子图发现的任务是在一个图的集合中找到感兴趣的或频繁出现的子图。⑥图分类是一种试图将整张图用正或负的标签来分类的监督学习问题，这是最早将机器学习和数据挖掘技术应用于图数据的任务。

（3）面向复杂数据类型的数据挖掘

除了上述两种（简单）数据类型之外，事实上，数据记录的手段众多，比如文本、图形、图像、音频、视频等，这也就意味着：数据挖掘（技术）应该对所有这些数据（类型）都有应对方案。因此在数据挖掘技术的分类中，专门有一类被称为“复杂数据类型挖掘”的技术（研究方向），专门针对上述诸如文本、图形、图像、音频、视频等数据类型进行分析和挖掘。不过针对这样的复杂数据类型，研究者通常的研究思路和做法是：通过特征提取的方法将有助于对这些数据对象进行分析的有效特征（向量）以“属性-值”的模式提取出来，然后利用传统的“属性-值”数据挖掘方法加以响应。因此，对于这类复杂数据类型的数据挖掘，更重要的问题往往是如何有效地提取特征，这就涉及对这些数据的理解问题，这往往需要专门的领域知识和相关技术支持的响应，因而也就引发出很多独立的研究方向，比如自然语言理解、多媒体处理等。

以图像理解为例，为了对图像进行有效的后续分析，必须对图像有所理解，而一般的图像存储格式为一个二维数组，如何从这个二维数组中提取出对未来分析有贡献的信息，也就是特征，至关重要，而事实上，就有专门的研究方向专注于做图像特征的表示与提取的研究，此处不赘述。一般的研究思路是：描述一个特征的三个重要（物理层）特征维度包括颜色、形状和纹理及其他一些高级语义特征，比如空间语义（图像中的两个物体A与B的位置关系）、情感语义（图像表示或反映的情感倾向如何）、属性语义（图像表示中的物理实体是什么）等。而其中最为关键的就是三大物理特征的提取（很多研究者关于后续高级语义特征的提取事实上都是在物理特征的基础上进行的），因而就产生了用颜色直方图表示图像的颜色特征；用不同的“矩”特征表示图像的形状特征；用共生矩阵或谱（基于信号处理方法）表示图像的纹理特征等，具体方法的执行及更多的相关算法此处不一一赘述。总之，通过这样的特征提取，一个图像就被转换成为一个特征矩阵，而此特征矩阵就是这个数据对象的属性特征，再往后，就演变成成为传统的“属性-值”数据挖掘的问题了。

（4）面向“时序模式数据”的时序数据挖掘

时序数据是一种有别于上述数据类型的数据，所谓时序数据广义上是指所有与时间相关，或者说含有时间信息的数据，但在具体的应用中，时序数据往往是指用数字或符号表示的时间序列。由于大多数场景下的数据产生都是有时间先后关系的，因此，数据的时序性几乎是数据的普适特性。时序数据是随着时间连续变化的数据（序列），因而其反映的是某个待观察过程中的状态或表现。由于时序数据一般都以时间为基准呈序列状排列，因而，对时序数据的挖掘也可以看作是一种比较特殊的序列数据挖掘，其主要目的仍然是描述与预测，前者是学习待观察过程过去的行为特征，比如顾客的消费习惯等；后者是预测未来该过程的可能状态或表现，比如可能会购买与历史上属性类似的顾客类似的货物。时序数据挖掘的主要研究内容包括数据预处理、时序数据表示、分割、相似度度量、分类、聚类、异常检测、（关联）规则挖掘等，此处不再一一赘述。

以上是从数据及数据对象的特点对数据挖掘的几个主要分类进行一个简单的介绍,事实上,不论是在哪一个分类归属上,数据挖掘的主要功能都是描述和预测,前者主要是为了了解数据中潜在的规律,往往通过一些具体的方法来进行数据建模,主要的方法有分类、聚类、关联规则挖掘等(前文已有专门的介绍,此处不再赘述);后者主要是利用这些模型在具体应用场景下进行使用,从而达到预测未来的目的(尽管这一步实际操作起来很难)。

在很多教科书中,异常检测(Anomaly Detection)也被归类到数据挖掘领域中。所谓异常检测指的是在数据中发现与预期行为模式不符的数据记录(这些与预期行为不符的数据记录被称为异常)。异常检测问题的研究来源于各行各业的实际应用需求,例如信用卡欺诈的检测、电信行业的违规检测、医疗保险的风险检测、网络安全中入侵行为的检测、系统失效检测、敌情行为监测等,因此异常检测可以看作数据挖掘中的一个应用问题。

一般异常检测的技术思路有3个(包括但不限于):①最直接的异常检测算法就是设计描述正常行为的规则,凡是违背规则的就认为是异常,显然,如何构建正常行为的规则,这本身就是极具挑战的问题。②将异常检测问题建模为聚类问题,只不过这种“聚类”并不关注每个“聚类”,而是关心那些远离各个聚类中心的数据对象(这些数据对象往往被称为“离群点”,因此,此类问题往往也被称为离群点发现),显然,度量的设计非常重要,而这也是极具挑战的。③将异常检测建模为一种特殊的分类问题,这也是当今异常检测领域的研究热点。如果应用的需求只是关注数据是否为异常,则可以当作二分类问题处理,如果应用的需求不只是关注数据是否异常,还关注异常数据的具体类型,则可以当作多分类问题来处理。但是这种基于分类模型的做法偏重于针对刻画正常数据的优化,而没有针对异常检测这个直接目标进行优化,这会使这类异常检测模型应用时或者产生太多的假警报(将原本正常的数据标记为异常数据)或者只能检测到少数异常(即假正常,将原本异常的数据标记为正常)。而假警报和假正常也是相互冲突的两个问题,即如果试图减少其中一种问题,另外一种问题则会随之变多。

或许正是因为异常检测在实际应用中有极大的需求,而且其中也隐含了甚多的技术难题,所以异常检测问题事实上是机器学习与数据挖掘研究者普遍关注的一个问题,并被上升为一个独立的研究方向。

数据挖掘的一般流程如图9-1所示。

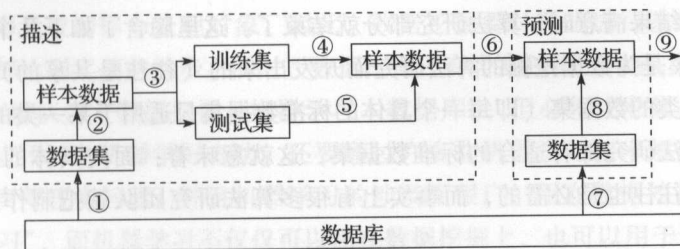


图9-1 数据挖掘的一般流程

如图9-1所示,数据挖掘的基本框架是在数据库意义上进行两件事情:描述和预测,一般而言,完整的数据挖掘流程包括的几个关键环节如下:

1) 步骤①的目标是从数据库中提取有效的数据集以用于后续的数据建模,有两点值得注意:①这里的数据库是一种广义的说法,可以是SQL、NoSQL、NewSQL甚至是文件,表示的是数据源。②这一步往往是实际操作中的一个关键节点,因为能否高效快速地从数据库中提取用于建模训练的数据直接关系到整个训练建模的效率,而数据库的并发访问受限,往往使得这个环节(仅仅读取)的时间耗费会很大。

2) 步骤②的目标是对提取的数据集进行必要的预处理,以便形成高质量的训练样本供后续的建模训练使用,主要内容包括数据清理(消除或减少噪声,处理空缺值,以减少学习时的混乱)、相关性分析(数据中的有些属性可能与当前任务不相关;也有些属性可能是冗余的;删除这些属性可以加快学习步骤,使学习结果更精确)、数据规范化(将数据概化到较高层概念,或将数据进行规范化)等,显然,这一步是极其重要的,因为如果得不到高质量的训练样本,后续建模训练的质量就得不到保障。

3) 步骤③的目标就是以不同的策略将样本数据分割为训练集和测试集,前者用于数据建模(步骤④),后者用于评估建模的效果(步骤⑤)。这3个步骤往往是耦合在一起且经常迭代的,比如根据不同的策略进行训练集和测试集的划分,然后进行训练,再然后进行评估,值得注意的是在步骤④中,往往会涉及特征提取与特征选择等问题,这一块在前文已经有所介绍,本章不再赘述。

4) 步骤⑥的目标是将训练完成(并且得到认可)的模型输出以便进行未来的预测。

5) 步骤⑦、步骤⑧、步骤⑨的目标是利用训练好的模型进行预测,具体而言,就是将数据库中的数据以雷同步骤①和②的方式进行提取,并以步骤④的方式进行特征提取与选择,然后将数据输入步骤⑥所输出的目标模型得到输出(预测结果)。

上述的流程是一个典型的面向具体应用场景的数据挖掘流程,但是在实际工作中,数据挖掘技术的研究往往是先于数据挖掘技术的应用的。这就意味着,对于数据挖掘技术的研究者而言,在进行研究的时候是没有数据库支撑的。因此,所有的数据挖掘研究者都必须依赖标准数据集,具体流程类似于上述标注流程中的第①步至第⑤步。也就是说,数据挖掘的研究者从标准数据集中提取数据、预处理数据(由于标准数据集的价值就是为了数据挖掘算法预研的,因此这一块往往已经做好了),然后进行训练集和测试集的划分,最后进行训练建模和评估,当对评估结果满意时,算法研究部分就结束了。这里隐含了如下几个问题:

1) 标准数据集是为数据挖掘的算法研究而开发出来的(往往是共享的),根据不同的研究目的,有不同种类的数据集(即每一个具体的标准数据集只适用于某一类的算法研究),但并不是说所有的算法研究都有适合的标准数据集,这就意味着:面向具体的研究目标开发有针对性的数据集,往往也是必需的,而事实上有很多算法研究团队就把制作数据集作为研究的一部分。

2) 所有的标准数据集都是为算法研究做基础的,但是标准数据集往往与实际应用场景中

的数据样本有着天壤之别,这就意味着在标准数据集上效果很好的算法在实际应用中往往表现得并不是那么优秀,甚至还会很糟糕。也就是说,在具体的应用场景下,以某一个算法为基础进行算法的改进是必需的。

3) 即便是利用实际数据进行建模能得到优秀表现效果的算法(上述步骤①至步骤⑤)在实际预测环节(上述步骤⑦至步骤⑨)也未必能够得到如同训练测试评估中的性能,一个重要的原因是建模所利用的数据是历史数据,而实际应用中面对的数据是不断新生的数据,而新生数据的模式和历史数据的模式是否一致是无法确认的。

尽管如此,数据挖掘算法的预研及实际应用研发还是齐头并进:不同的组织不断开发的标准数据集及一些基本数据挖掘算法、工具及平台让算法的研究者有更多的精力专注于算法本身;而不断进展的算法研究成果也给应用研究者以更加专业的算法储备,应用开发研究者在面向具体的应用场景下对算法进行不断的改进和迭代,或许这本身就是应用研究者的价值所在。

9.2.5 应用提示

作为人工智能的一个分支,数据挖掘无疑是数据驱动的一个研究热点,特别是在大数据如此炙手可热的当代。数据挖掘是使用一类实用的应用算法(大多是机器学习算法),利用各个领域产出的数据来解决与该领域相关的问题,求解过程需要用到不同研究领域的不同技术。对数据挖掘而言,3个重要的技术支撑分别是数据库技术、机器学习和统计学,其中数据库提供数据管理技术(本节不再赘述),机器学习和统计学提供数据分析技术。某种意义上而言,统计学、机器学习与数据挖掘都可看作数据分析的工具,但是三者之间有着很大的区别,至少包括(但不限于):

1) 机器学习是一门涉及自学习算法研究的科学,统计学(大多是推断统计学)往往是研究的基础和工具,机器学习研究的算法本质上是通用的,可以应用到不同的应用领域(数据挖掘仅是其中的一个)。

2) 统计学往往专注于理论的优美而忽视实际的效用,因此,统计学界提供的很多技术通常都要在机器学习界进行进一步的研究,变成有效的机器学习算法之后才能再进入数据挖掘的领域,可以认为统计学是通过机器学习来对数据挖掘产生贡献的。

3) 从数据分析的角度来看,绝大多数数据挖掘技术都来自于机器学习领域,但机器学习研究往往并不把海量数据作为处理对象,因此,数据挖掘要对算法进行改造,使得算法在性能和空间占用上都达到实用的地步。同时,数据挖掘还有自身独特的内容,即关联分析。

4) 可以粗放地认为“数据挖掘=机器学习+数据库”,但必须注意到:数据挖掘不仅仅要研究、拓展、应用一些机器学习方法,还要通过许多非机器学习技术解决数据仓储、大规模数据、数据噪声等现实问题;并且机器学习的涉及面很广,用在数据挖掘上的方法通常只是“从数据中学习”,而机器学习不仅仅可以用在数据挖掘上,也可以用于诸如自动控制等这些与数据挖掘几乎无关的领域。

5) 统计学习更倾向于模型,通常会基于某种已知的模型进行计算,因而对模型的要求比较苛刻,统计建模是基于数据的概率来分布的,因此统计模型很重视推断,这些推断,比如假设检验、置信区间等都是基于某种分布假设的,并且认为这个世界是可以概率分布来逼近的。而机器学习并不在乎数据产生于什么分布,并且认为这个世界运行的方式是无法单纯用概率分布来解释的,因此,它的目的是预测的精准性(最小化预测误差的某种度量)。

9.3 从数据挖掘到数据科学

9.3.1 从“惊奇”引发的科学之母

哲学(Philosophy)源自希腊语“Philosophia”,“Philosophia”是由 Philo 和 Sophia 两部分构成的动宾词组:前者指爱 and 追求;后者指智慧。因此哲学的本意是爱智慧,即人类为了更有智慧而进行的思想认识活动。1874年,日本启蒙家西周在《百一新论》中将“Philosophy”用汉文翻译为“哲学”,1896年前后康有为等将“哲学”的翻译介绍到中国,后渐渐通行。

在原始社会中,人类对各种自然现象还不了解,打雷闪电、山洪暴发等自然现象激起了人类对自然和自身的探索 and 认识,这便是宗教的早期雏形,可以说,在这一时期,哲学以宗教的形式存在。人类进入古代奴隶制时期后,社会经济的发展提高了人类的认识能力,人类开始思索世界的本质等理论问题,人类早期的哲学思想开始出现了。

事实上,关于哲学的定义,不同的学者从不同的视角有不同的认知。在很多学者的眼中,哲学是介乎神学与科学之间的东西。英国哲学家罗素就认为,“一切确切的知识都属于科学,一切涉及超乎确切知识之外的教条都属于神学,介乎神学与科学之间还有一片受到双方攻击的无人之域,这片无人之域就是哲学。”因此,一方面哲学和神学一样,包含着人类对于那些迄今仍为科学知识所不能肯定之事物的思考;同时哲学又像科学一样是诉之于人类的理性而不是诉之于权威。黑格尔认为哲学是对绝对的追求,他在《小逻辑》中提到“哲学是以绝对为对象的特殊思维方式的思维运动”;冯友兰在《中国哲学简史》中提及“哲学就是对于人生的有系统的反思思想”;胡适在《中国哲学史大纲》中指出,“凡研究人生且要的问题,从根本上着想,要寻求一个且要的解决”这样的学问叫作哲学。

一个共识是——中外哲学的产生皆起源于疑问。柏拉图指出“thauma”(惊奇)是哲学家的标志,它是哲学的开端;亚里士多德在《形而上学》中说:求知是所有人的本性,人都是由于惊奇而开始哲学思维的,一开始是对身边不解的东西感到惊奇,继而逐步前进,而对更重大的事情发生疑问。

爱因斯坦关于哲学的定义是:如果把哲学理解为在最普遍和最广泛的形式中对知识的追求,那么,哲学显然就可以被认为是全部科学之母。在古代,哲学研究的对象是庞杂的,凡是能给人以智慧、使人聪明的各种问题,都是哲学研究的对象。这时期的哲学研究对象,包含了具体科学的对象,哲学和科学浑然一体。到了奴隶社会中期,数学、天文学和医学等具

体科学成为一门独立的科学，从哲学中分化出去了；随着资本主义社会的确立，产生了近代实证科学，各门具体科学纷纷从哲学中独立出去，获得了突飞猛进的发展；在当代，由于自然科学、社会科学和意识科学的独立和迅速发展，哲学的研究对象又发生了变化。哲学不再研究世界某一范围、领域的问题了，而是研究整个世界一切事物、现象的共同本质和普遍的规律，如世界的本源，物质和意识的关系，世界的基本状态等问题，从而形成了唯物唯心两大派系。唯物主义将世界的本源归结为物质，主张物质第一性，意识第二性，具体又分为古代朴素唯物主义、近代形而上学唯物主义与辩证唯物主义；唯心主义将世界的本源归结为精神，主张意识第一性，物质第二性，物质是意识的产物，具体又可分为可知论、不可知论、二元论、知识论等。

科学一词，来源于拉丁文“Scientia”，意为“知识”、“学问”，在近代侧重于关于自然的学问，对应的英文是“Science”，日本启蒙思想家西周将其用汉文翻译为“科学”。到了1893年，康有为引进并使用“科学”二字。严复在翻译《天演论》等科学著作时，也用“科学”二字，从那时开始，中国人使用“科学”一词的频率逐渐增多，并且一直延续至今。

科学是关于自然界、人类社会和人自身规律的事实、原理、方法和观念的知识体系，以及创建这个知识体系的社会活动。科学的任务是发现规律、提出理论、认识世界、解释世界，通常认为，科学至少包括以下3个方面的内容（但不限于）：

1) 科学是人们研究自然、社会、思维的本质及其规律所获得的一种知识体系，因此科学知识是从科学实践中抽象出来，又为科学实践所证实的，反映客观事物的本质和规律的理论体系，具有客观真理性、普遍性、系统性、逻辑性、解释性，不属于上层建筑，没有国界、阶级性、民族性，是一种知识形态的生产力。

2) 科学不仅是一种知识体系，它还是产生知识体系的一个活动、一个过程，科学（包括技术）归属于劳动（人类使用工具来认识世界、改造世界，以满足生存需要的各类活动）领域，并构成了劳动领域的主体与特征。

3) 随着科学日益渗透到社会的各个方面，科学已经成长成为一种重要的社会事业，科学作为一种社会建制反映了其作为社会系统的一个重要部分的结构特性，同时科学是一种集体的事业，随着社会职业和社会部门被越来越细致地分类，科学在社会中越来越结构化，与社会的关系也越来越紧密。

笛卡儿（R. Descartes）被认为是现代科学的鼻祖，笛卡儿从“我思故我在”的概念出发，认为思维是人存在的基础，但强调了感觉的重要性：只有通过感觉才可以思维，但是人的感觉是不可靠的。因此，笛卡儿认为只有通过通过对自然现象的观察、分析，并经过所谓的“逻辑推理”得出的结论才是可靠的、唯一的、泛普的，才有可能成为知识的一部分。但是他同时指出了知识的有限性，进而提出所谓的“质疑说”。对未知真理的逼近，就是对现有知识的不断质疑和对未知的探索，于是，质疑便成为向未知推进的原动力。笛卡儿认为真理的发现可以有其他独立的途径（比如通过数学），而不是通过信仰，他在质疑（挑战）当时真理的权威

(基督教)时提到:人类的意愿就是上帝的意愿,所以人类完全可以独立于上帝。笛卡儿的类似于上述的认知现在看来已经是不争的事实,但是在17世纪,这种认知完全是思维的革命、科学的开始。

对于(自然)科学,求解或寻求真理的途径完全依赖于假设、证明、推理和实验,这种探索的途径与方式和信仰什么宗教、运用什么语言、基于何种文化等均无关。研究者仅需要应用数学语言和科学的方法达到求证的目的,而没有必要向对方做任何其他的解释(因为科学意义下的真理是唯一的、可重复的、泛普的,这与科学之外的领域需要通过辩论、争论、说服甚至对抗才可以定论的“真理”完全不同)。也就是说,科学与信仰(belief)无关、与感觉(sense)无关、与看法(perception)无关,因为信仰、感觉和看法等是主观的、因人而异的,而科学是唯一的、客观的。

对于科学的特征,不同的时期及不同的学者一般都有不完全一致的认同,基本的认同包括(但不限于):①理性客观。一切以客观事实的观察为基础,通常科学家会设计实验并控制各种变因来保证实验的准确性及解释理论的能力。②可证伪。由于人类无法知道一门学问里的理论是否一定正确,但若这门学问有部分错误时,人们可以严谨明确地证明这部分的错误,那这门学问就算是合乎科学的学问了。③存在一个适用范围。任何一门科学都是有适用范围的,而不是放之四海皆准的,尽管有科学家仍然在努力寻找与探索是否有某种理论可以囊括所有的自然现象。④普遍必然性。科学理论来自于实践,也必须回到实践,它必须能够解释其适用范围内的已知的所有事实,同时具有可重复性。

特别值得一提的是:自人类文明以来,在人类认识自然和改造自然的过程中,产生了很多截然不同的方法和理论(用于探索自然),笛卡儿提出的科学哲学及方法论而引发的科学(活动)仅是其中的一种,因而,我们不能把人类的所有理论与方法都归结于科学之中,如果是这样盲目地归类,本身就不科学。

习惯上,人们倾向于将科学和技术联系在一起,并用另外一个词“科技”来表示,而实际上,科学和技术这二者既有密切联系,又有重要区别。

科学是认识世界的工具,其目的是提高人类的认知水平,科学发展的动力主要是科学家的好奇心、兴趣和社会责任感,尝试发现自然界中确凿的事实与现象之间的关系,并建立理论把事实与现象联系起来,由于科学的研究往往是和未知的领域打交道,其进展,尤其是重大的突破,是难以预料的;而技术的任务则是把科学的成果应用到实际问题中去,以增强人类的生存能力、改善人类的生活质量。古代的技术主要来自于生产实践,现代技术则更多的是根据一定的科学原理,为达到一定的应用目的,所发展和开发出来的方法和手段。

在实际的社会分工中,科学与技术研究可分为基础研究、应用研究和技术开发:基础研究没有特定的商业目标,是以探索规律、发展原理、提出理论为目标的科学发现活动;应用研究是以工程为目标,探索新知识应用的可能性(关注的是技术的泛化应用性能);技术开发(往往也称为项目开发、工程开发等)是把科研成果应用到生产工程上的技术创新活动,比如

产品开发、设备与工具开发、生产工艺开发、能源和原材料开发、改善生产环境开发等。

综上所述,哲学是凌驾于几乎所有学科研究之上的科学之母,关注的是意识形态意义上的思维方式,形而上学、逻辑学、认识论、伦理学及美学等是哲学的几个主要学科分支;而科学关注的是(自然)科学规律的发现,一般基于观察、假设、验证的思路进行,往往与伦理、价值等无关;技术关注的是利用既有的科学成果实现某种功能,至于内在的道理并不关心(那是自然科学需要解决的事),技术的设计者与实现者大多能使得自己的思路暗含科学道理,但是他们不见得要从根本道理上去揭示,只要功能满足要求就可以了,技术的价值体现在其泛化的应用及性能方面;工程的目标就是造出某个东西来满足具体领域的目标需求,价值体现在工程项目成果的使用带来了多少收益。

计算机学科是围绕计算展开的一门学科,因此,在计算机这个领域,计算机科学关注的是“Whether to Compute?”(是否可计算),即计算机科学关注目标问题是否可计算;计算机技术关注的是“How to Compute?”(如何计算)即用什么样的手段对可计算的问题进行计算;计算机工程关注的是“What to Compute?”(在哪个领域应用),即针对什么样的领域需求展开各类计算。

9.3.2 从“科学”引发的研究范式

范式(Paradigm)指的是从事某一科学的研究者群体对本体论、认识论和方法论的基本承诺,是科学家们所共同接受的一组假说、理论、准则和方法的总和,这些东西形成科学家心理上的共同信念。范式这个概念和理论是由托马斯·库恩(Thomas Kuhn)提出的,并在1962年出版的《科学革命的结构》(The Structure of Scientific Revolutions)中进行了系统的阐述,因此“范式”这个概念本意上是哲学意义上的思维方式,也因为不同的研究领域而衍生出了若干种范式类型。

由于“范式”这个词在语义上的“普适性”,几乎可以套用到传统与创新的任何领域,使得“范式”的使用过于“弹性”。比如工业人士认为“蒸汽机的发明”引发了产业结构的革命(范式的变化)使得人类进入工业1.0时代,因为“电力的发明”引发了产业结构的革命(范式的变化)使得人类进入工业2.0时代,因为“信息技术的发展”引发了产业结构的革命(范式的变化)使得人类进入工业3.0时代,继而因为“人机融合的理念”引发了产业结构的革命(范式的变化)使得人类进入工业4.0时代,也是现在我们所正处于的时代。经济学家则运用税收政策来促进公司结构(范式)的合理化;管理学中则出现了一系列诸如组织范式、开放范式、同步范式、协同范式、参照范式和随机范式等新术语;社会学家则利用“范式”来描述“社会范式”的变化,等等。

本章的行文是从哲学的视角来看待“范式”这个概念的。2007年1月11日,吉姆·格雷在NRC-CSTB(National Research Council-Computer Science and Telecommunications Board)大会上的演讲“科学方法的革命”中提及:所谓范式,就是某种必须遵循的规范或大家都在用的套

路,科学研究范式可分为实验归纳、模型推演、仿真模拟和数据密集型科学发现(Data-Intensive Scientific Discovery)。其中,最后的“数据密集型”就是我们现在所说的“科学大数据”。

(1) 实验归纳

“实验归纳”(第一范式)是人类最早的科学研究方式(现在仍然是很多研究领域的主要研究方式),主要以记录和描述自然现象为特征,在研究方法上,以归纳为主,带有较多盲目性地观测和实验。这种方法自从17世纪的科学家弗朗西斯·培根(Francis Bacon)阐明之后,科学界一直沿用至今。弗朗西斯·培根指出科学必须是实验的、归纳的,一切真理都必须以大量确凿的事实材料为依据,并提出寻找因果联系的科学归纳法。

(2) 模型推演

被称为“第二范式”的模型推演偏重于理论总结和理性概括,强调普遍的理论认识而非直接实用意义的科学。在研究方法上,以演绎法为主,不局限于描述经验事实。该研究范式强调数据模型的构建,以超凡的头脑思考和复杂的计算来超越实验设计(第一范式注重的),而随着验证理论的难度和经济投入越来越高,科学研究的难度也越来越大。

(3) 仿真模拟

20世纪中叶,由于计算机的发明及不断发展,利用计算机对科学实验进行模拟仿真的模式得到迅速普及,人们可以通过对复杂现象进行模拟仿真,推演出越来越多、越来越复杂的现象,涉及的问题主要有数值模拟、模拟拟合与数据分析、计算优化等。随着计算机仿真模拟越来越多地取代实验,从而逐渐成为科学发现的第三种主流思路,即第三范式。第三范式本质上是假设驱动的,即先提出可能的理论,再收集数据,然后通过计算来验证。在这个研究范式意义下,计算机被当作是一种工具,其优势是“高速计算”。

(4) 数据密集型科学发现

随着计算机计算性能的不断提高,在进一步彰显“仿真模拟”优势的同时,数据得到不断采集,并成为科学研究中越来越不可避免和必须响应的事实。基于此,人们提出的一个科研假设是:能否从传统的假设驱动转向以数据驱动进行科学研究,这就引发了被冠以“第四范式”标签的数据密集型科学发现思维,人们希望通过对已经有(并不断地、有意识地采集)的大量已知数据的有效分析和计算,得出未知的理论。或许这种思路既不能像实验(第一范式)那样明确地告诉你“是什么”,也不能像理论(第二范式)和模拟(第三范式)那样在一定程度上告诉你“为什么”,但是,“第四范式”可以告诉你“大概是什么”,而这种“大概”是隐藏在数据背后的“客观”,让计算机自己从海量的数据中发现模式,也就是共性、客观,这本身就是科学发现所需要的。

事实上,上述的4类科学发现研究范式是进行科学研究中的4种思维方式,不是说哪一种方式可以取代另外一种方式,而是我们在实际的科研工作中,应当根据实际研究领域的特点,在思维及方法论层次选择需要遵循的方式。当然,也可以作为第三方,根据具体依据的“范式”来区分不同的科学家(工作者)所处的角色群体(共同体或亚共同体)。

从科学研究的角度来看,基于密集型科学发现已经成为科学研究者的一个共识,这也是

在大数据概念不断得到热议和认可的当口,很多学科都自发地产生“大数据+”研究方向的原因。在大数据时代,科学家不仅要通过对广泛的数据实时、动态地监测与分析来解决难以解决的科学问题,还要把大数据作为科学研究的对象和工具(而不是像第三范式意义下的仅将计算机作为工具),基于大数据来思考、设计及实施科学研究。

事实上,即便是在大数据应用场景下,一般大数据项目的建设也有一个隐含的需求:通过大数据平台(含计算平台)的建设为领域专家提供领域分析并为领域研究提供各类支撑服务。其所追求的就是从领域应用的角度,利用大数据发现领域的知识和规律。

但是应该注意到:将数据作为研究的驱动力是不同领域科学家所依赖的研究范型,而为不同学科的科学家提供数据驱动的研究动力,则是大数据工作者的分内工作。

9.3.3 从“数据”引发的数据科学

“密集型数据”成为科学发现的第四范式,并得到各个领域学科研究者的共识,表明了研究者从科学研究的角度对“大数据”能力的认可和期待,从为科学研究服务的角度而言,大数据是一个工具或技术。就这个意义而言,大数据是一个面向目标应用服务的应用(软件、系统、平台等),大数据的“价值”体现在可为科学研究提供支撑。

随着大数据概念得到“政产学研商用”各界的普遍认同,在这个风口下已经建立起或正在建立的若干大数据平台,都在不断暗示各界对大数据价值的认同,大家都认为从大数据背后发现的模式能给大数据项目的建设者带来更多的商业获益,在这个意义上,大数据的“价值”体现在为项目建设主体带来实际的增益。

大数据作为一个“媒介”将各边力量联系在了一起,并在为了共同目标进行协作的过程中所彰显的“集体智慧涌现”也从另外一个方面彰显了大数据的“价值”。

层出不穷的大数据项目建设,以及大家主观共识的大数据价值都在暗示:大数据是一个应用问题。基于价值驱动的理念,可以粗放地认为:没有价值期望的大数据项目是不可能落地的。

而在大数据的建设过程中,数据的4V特征给传统的数据采集、存储、分析、系统实现都带来了前所未有的若干挑战,这些挑战最终可归结为:如何进行有效的数据存储?如何保证数据的高并发访问?如何进行有效的数据分析?如何设计、实现大数据平台等。所有这些事情上都是每一个大数据应用实施者必须面临的关键问题,这也意味着,大数据是一个技术问题。既然是技术问题,针对具体的目标需求,如何提高技术应用的普适性及泛化性,应当是大数据技术研究者需要关注的重要话题。

或许更重要的问题在于:大数据是一个科学问题。

计算的本质是进行诸如 $y = H(x)$ 的计算,其中, $H(\cdot)$ 就是算法或软件(程序), x 就是输入和数据,计算 y 的本质就是一个程序基于给定的输入,利用某个算法处理后,输出结果。

传统的计算机科学研究几乎都是围绕这个算法展开的,并以这个算法为研究基础的(通过判断算法是否可计算,来决定计算机能不能处理)。传统的研究是假设输入 x 不变(不重

要), 主要的研究集中在 $H(\cdot)$, 在此基础上研究各类图灵计算意义下的若干算法 (多项式算法、近似算法、随机算法等), 即便如此, 算法的研究也几乎是举步维艰。而在大数据场景下, 问题进一步演变为输入 x 也是变化的, 那么在大数据场景下传统的计算理论是否还有效?

另一方面, 传统的研究一般是基于封闭的假设, 即数据都齐全了以后再进行计算, 而在大数据场景下, 数据是增量到来的, 这意味着每次针对数据的计算都是不完备的, 如何利用历史上不完备的输入进行计算并综合实时到来的新增数据进行增量计算, 并充分利用两者的计算实现完整数据的完整计算功能, 这势必会涉及计算理论的进一步研究。

第三, 传统的计算都是在一定的数据输入基础上, 利用既有的公理系统 (定理系统) 进行演绎计算的, 而在大数据场景下, 输入计算的数据是有相关性的, 这种相关性的发现需要归纳, 如何有效地将归纳融入演绎的过程中也是计算理论需要解决的问题。

大数据给传统理论带来的上述挑战意味着必须将大数据建模为一门计算问题, 从“算法 + 数据”的角度进行研究, 这是典型的科学问题, 针对的是可计算问题。

或许正如李国杰院士在提及数据科学及其研究对象时说到的那样: “计算机科学是关于算法的科学, 数据科学是关于数据的科学”。同时他提及“数据背后是网络, 网络背后是人, 研究网络数据实际上是研究人组成的社会网络”。因此, 大数据已成为联系人类社会、物理世界和信息空间的纽带, 需要构建融合人、机、物三元世界统一的信息系统。

需要进一步考虑的是: 大数据或许本质上是一个哲学问题。比如: 大数据是什么? 大数据从哪里来? 到哪里去? 可以纯粹地从技术应用的视角来说大数据就是“大”的数据, 大数据从数据采集而来, 应用 (服务) 于具体的目标应用。但是上升到逻辑意义或更高层的一些问题, 比如“大数据的本质是什么”、“我们真的需要大数据吗”、“我们应该如何拥抱大数据”等, 慎思并回答上述问题已经不是科学、技术及应用层面上事了。澳门大学校长赵伟提及的“数据科学是介于哲学和自然科学之间的学科”, 或许出发点就在于此。

9.4 从算法到大数据方法论

9.4.1 演绎与归纳

归纳逻辑是研究归纳法、归纳推理的逻辑, 与归纳逻辑相对应的, 演绎逻辑是研究演绎法、演绎推理的逻辑。归纳及归纳推理关注的是从个体推出一般, 即以一系列经验事物或知识素材为依据, 找出其服从的基本规律或共同规律, 并假设同类事物中的其他事物也服从这些规律, 从而将这些规律作为预测同类其他事物基本原理的一种认知方法; 而演绎及演绎推理关注的是从一般推出个别, 也常被称为必然性推理, 或保真性推理。

根据“一切人都会死”和“苏格拉底是人”这两个陈述, 我们可以很容易地得出“苏格拉底也会死”的结论, 这里使用的推理就是演绎推理。从“燕子是卵生的”“麻雀是卵生的”

“大雁是卵生的”“老鹰是卵生的”“常见的鸟仅包括燕子、麻雀、大雁、老鹰”这几个陈述，我们可以很容易地得出“常见的鸟都是卵生”的结论，这里使用的推理就是归纳推理。

对于人类认识、实践活动而言，归纳和演绎都是不可或缺的推理方式，更让人不可思议的是：系统的演绎逻辑理论是两千多年前的亚里士多德创始的，而古典的归纳逻辑是由迟于亚里士多德近两千年的弗兰西斯·培根创立的。

演绎推理常遭诟病的一点是：演绎的逻辑功能只是从给予的陈述中把真理传递到别的陈述上去，而这是它所能办到的全部事情了，换句话说：演绎推理是空虚的，结论并不能陈述多于前提中所说的东西，它只是把前提中蕴涵着的某种结论予以说明而已。

1843年，穆勒（John Stuart Mill）在《逻辑、推理和归纳体系》一书中提到：“科学与考察是从对事实的自由、无偏见的观察开始的，接着又对这些事实进行归纳推理而形成一般规律的公式，最后进一步归纳到更广的一般性，形成人们所称的理论，最终又要把规律和理论的经验结果同所有观察过的事实，包括最初开始观察的事实进行比较，来核对规律和理论的真实内容”。这样，由逻辑实证主义发展起来的“假设→演绎”就是以这种归纳逻辑为基础的，“假设→演绎”的前提和新知识的获得都来源于对经验事实的观察和实验。演绎逻辑与归纳逻辑的结合就形成了一种“归纳→演绎→再归纳检验”的方法论。由此可见，按照经验论的观点，假设来源于对经验事实的归纳，知识的发现也依赖于对经验事实的再归纳，逻辑演绎对此则毫无用处。

从19世纪末到21世纪初的百余年中，归纳逻辑经历了从古典类型向现代类型的演化，并获得了长足的发展。现代归纳逻辑的发展及其在博弈、决策、计算机与人工智能、法学、认识论与方法论研究中的广泛应用，使其成为一门具有广阔前景的逻辑理论。归纳逻辑的发展历史大致可以分为三个阶段：①第一阶段是从17世纪20年代到19世纪中叶，称为古典归纳逻辑阶段，代表人物是穆勒。②第二阶段是从19世纪中叶到20世纪20年代，称为古典归纳逻辑向现代归纳逻辑过渡的阶段，代表人物是凯恩斯（John Maynard Keynes），其创建了第一个现代归纳逻辑的系统理论，将概率看作两命题或命题集合之间的一种逻辑关系，由此建立了第一个公理化的概率演算系统（是否考虑归纳前提与结论间的概然性是古典归纳逻辑与现代归纳逻辑的根本区别）。③第三阶段是从20世纪20年代至今，称为现代归纳逻辑的发展阶段，各种类型的归纳理论相继问世并得到不同程度的发展，比如概率理论，以及基于模态逻辑及模型论的现代逻辑理论等，代表性人物有莱辛巴赫（H. Reichenbach）、卡尔纳普（R. Carnap）、伯克斯（A. W. Burks）、科恩（L. J. Cohen）、冯赖特（G. H. von Wright）等。

其实，归纳逻辑的思维方式是机器学习与数据挖掘的基本方法论：利用训练集（经验事实）进行建模，获得模型（新知识），然后利用测试集进行测试（再归纳检验）。

以分类为例，分类是找出“属性→类标”的模型，其中采用的方法就是机器学习中的所谓的分类算法，如决策树、神经网络、SVM等。分类算法本身并不关键，关键是所有算法均是从数据集中提取分类的模型，从而实现分类。模型的提取过程便是根据已有数据集进行归纳

的过程；关联分析是挖掘海量数据中符合特定支持度和置信度的关联规则，它根据已有数据统计得来，使用的也是归纳法；聚类分析是将数据按照相似程度划分为簇，与分类相比，可以称为非监督分类，使用的也是归纳法；异常检测是发现数据中的离群点，一种方法是通过发现数据集中的模型，从而寻找不能与模型完美匹配的点，模型是归纳得来的，检测是基于模型的，因此它也属于归纳方法；对于大数据模型的预测方面，根据模型进行预测，属于从一般到个别的论证，大数据模型具有一定的预测能力，但并不具有必然性，也属于归纳方法。

事实上，基于经验主义的归纳逻辑是存在问题的，即能否通过“是”推导出“应该”？或者“事实”命题能否推导出“价值”命题？或者对特殊的经验事实的归纳能否得出普遍的命题？这个问题是由休谟（David Hume）在《人性论》中提出的，并被标签化为“休谟问题”，这个问题在西方近代哲学史上占据了重要的位置，许多著名哲学家纷纷介入，但终未有效破解。

休谟在《人性论》中指出：归纳在逻辑上是难以成立的，因为没有什么逻辑容许我们确认“那些我们不曾经历过的事例类似我们经历过的事例”。例如，我们通常看到的天鹅都是白色的，但我们并不能由此得出天鹅都是白色的这一普遍命题或一般规律，或许遍历全世界所有的天鹅都是白色的，借此得到“天鹅都是白色”的结论或许是对的，可实际上我们无法“遍历”世界上所有的天鹅。休谟问题的本质问题在于：以过去的经验为基础推断将来要经历的任何东西并没有逻辑上的依据。因此，从对特殊事例的归纳到普遍的规律，在思想上就需要一种非逻辑的跳跃，其结果可能是由真的事例导出假的命题。

我们对于这个世界的认识，其实完全只是一种主观的判断（精神层面的，与人的价值观相关），这种判断和真实的“客观世界”是否一致，我们永远也不可能知道。虽然某些唯物主义者总喜欢用“无数次的实践”来证明主观与客观理论上最终能达到一致，但实际上，“无数次的实践”是不可能做到的，所以说这也只是一种空想罢了。

昔者庄周梦为蝴蝶，栩栩然蝴蝶也，自喻适志与，不知周也。俄然觉，则蓬蓬然周也。不知周之梦为蝴蝶与，蝴蝶之梦为周与？周与蝴蝶则必有分矣。此之谓物化。（摘自《庄子·齐物论》）

这个看似荒谬的故事显示了庄子不同凡俗的思维方式，以及其超越常人的精神与生命境界的思维。而其中隐含的一个哲学问题是：我们真的是我们认为的样子吗？这个问题在希拉里·普特南（Hilary Putnam）所著的《理性、真理与历史》一书中有所提及，并被标签化为“缸中之脑”问题，这个问题的原型如下：

一个人（比如是正在读这段文字的“你”）被邪恶的科学家施行了手术，他的脑被从身体上切了下来，放进一个盛有维持脑存活营养液的缸中。脑的神经末梢连接在计算机上，这台计算机按照程序向脑传送信息，以使他保持一切完全正常的幻觉。对于他来说，似乎人、物体、天空都还存在，自身的运动、身体感觉都还可以输入。这个脑还可以被输入或截取记忆（截取掉大脑手术的记忆，然后输入他可能经历的各种环境、日常生活）。他甚至可以输入代码，“感觉”到他自己正在这里阅读一段有趣而荒唐的文字。

有关这个假想的最基本的问题是：“你如何担保你自己不是在这种困境之中？”

事实上，我们真的没有办法证明我们不是“缸中之脑”，就像我们没有办法证明上帝存在或不存在那样。或许这些并不重要，我们并不需要去证明这件事情，或者试图去思考如何应对这件事。出于方便，我们可以假想背后有这么一个实体（比如“上帝”），这个实体仅仅是作为一个方便我们理解客观世界的模型，而并不需要真的存在。真实存在的那些就是：我们的感觉。作为一门工具，逻辑的价值在于应用，而不是将既有的算法（理论）静态地分类到某种概念框架的逻辑中，逻辑背后的思维方式才是大数据应用场景下数据分析师应该随时拥有的情怀。

9.4.2 因果与相关

因果性是指原因与结果之间的必然联系。因果性是贯穿西方哲学的一个极其重要的核心问题，也是至今西方哲学界所讨论和争论不休的一个热门话题。

之所以如此看重因果性范畴的研讨，一个重要原因或许在于将全部自然科学主要建立于因果性之上恰是（西方）科学精神的根本特色。比如亚里士多德在《分析后篇》中提到：“当我们认为自己认识到事实所依赖的原因，而这个原因乃是这事实的原因而不是别的事实的原因，并且认识到事实不能异于它原来的样子的时候，我们就认为自己获得了关于一件事物的完满的科学知识”。因此他致力于在一切事物中寻求“四因”，即质料因、致动因、形式因和目的因，这一科学传统自亚里士多德以来一直被延续下来。在中世纪和近代，对上帝存在的“存在论”证明、宇宙论证明和目的论的证明，实际上可归结为企图从亚里士多德的形式因、致动因和目的因来推出上帝的存在；近代实验自然科学则把原因大多理解为质料因和致动因（物质和运动）。总之，在西方哲学史和科学史上，人们通常认为任何现实事物都有其原因，而要把握一件事物的本质就必须找出它的原因。

大多数人都相信只要一事物伴随着另一事物而来，两事物之间必然存在着一种关联，使得后者伴随前者出现，并且因为强化了这种“必然”性而认为这两件有先后时序的事件是一个“因果”关系。对于大数据分析师而言，更有可能的判断或许是，这两件事情存在着“关联”（或然性的联系）而非“因果”（必然）。大数据分析师如此判断的出发点可能在于他们都拥有一个共同的认知：“要相关，不要因果”，究其思维的本源，或许休谟关于“因果问题”的主张能够暗合此思维方式。

休谟在《人性论》及后来的《人类理解论》一书中提到：虽然我们能观察到一事物伴随着另一事物而来，但是并不能观察到任何两事物之间的（因果）关联，只能得知某些事物总是“经常联结”（constant conjunction）在一起的。休谟提及的“经常联结”表示当我们看到某件事物总是“造成”另一事物时，所看到的其实是一事物总是与另一事物“经常联结”。因此，并没有理由相信一事物确实会造成另一事物，两事物在未来也不一定会一直“互相联结”。我们之所以相信因果关系并非因为因果关系是自然的本质，而是因为

们所养成的心理习惯和人性所造成的对这种“经常联结”在想象中的归类。

休谟主张人类（以及其他动物）都有一种信赖因果关系的本能，这种本能来自神经系统中所养成（并且无法移除）的习惯，但并没有任何论点、也不能以演绎或归纳的方式来证明这种习惯是正确的。由此可见，休谟对因果律的摧毁并不是丢弃了“因果性”这个名称，而是否定了它的客观必然性，把它变成了一种主观心理的习惯性联想。但这样一来，作为一种必然规律的“因果律”就不存在了。休谟之后的一些哲学家如伯特兰·罗素进一步强化了对因果关系理论的驳斥，甚至认为“因果关系理论是一种迷信”。

休谟对因果关系理论的驳斥（对经验归纳作为因果性的普遍必然基础的彻底否定）使当时所有坚持因果关系的哲学家们都束手无策，并使由因果律支撑起来的整个科学知识大厦面临崩溃的危机。康德作为一个理性的捍卫者从形式逻辑原则上升到“先验逻辑”层次，并从中梳理出包括因果性在内的一整套纯粹知性范畴，通过“先验演绎”证明了这些范畴运用于经验对象之上的必然性和普遍有效性从而达到对因果律的重建。康德对因果律的重建一方面认同了休谟基于经验主义的理解，即每件事的具体因果关系是后天得知的，并不能预先断言一件事的原因就是某件事；另一方面则坚持了理性主义对因果律的先天的普遍必然性和客观有效性的肯定，即主张任何“发生的事”都必定有其原因，尽管是何种原因还得诉诸于经验。

康德举例说：当我们观看一栋房子时，既可以从上到下，也可以从下到上，这种时间的相继只发生在我们的主观中，我们不会把这些局部的房子表象的相继呈现看作有什么客观的因果关系，它们的次序完全是可逆的，由此所形成的判断则只是“知觉判断”而不是“经验判断”。但现在如果有一件“发生的事”，如一只船顺流而下，我们一定是先看见它在上游，而后看见它到了下游，这两个表象在时间上的次序是不可逆的，这时我们就必须把主观中所看到的次序归之于客观上必然的因果关系了（“水流”是“小船从上游到下游顺流”的原因）。因此，凡是要认识一件客观上“发生的事”并由之形成经验判断，就不能不用到因果律，只有在这个先天条件下，任何发生的事才能够想象（和判断）。

康德认为因果律并不会因为具体某件事的原因不可预测，或是由于这个原因而不是那个原因而有所改变，它本身是绝对可靠的。这样一来，自然科学的基础由于被置于认识主体的先验范畴之上，因此就得到了不受经验偶然性所干扰的确立。明确了这一点，我们就可以放心大胆地去追究每一件事情发生的客观原因，并依据这一普遍必然性原理，一方面逐渐排除那些虚假的因果判断，另一方面日益揭示出更深层次的原因，而决不向超经验的“奇迹”之类的解释投降，也不归之于我们纯粹主观的心理状态。

康德关于因果性问题的研究可以归纳为如下两点：

1) 因果范畴只是解释一件已发生事情的必要条件，而不是充分条件，也就是说：

①因果范畴只需要解决“自然科学如何可能”的问题，说明如果没有这些先天的知性范畴，各种自然规律就失去了确定性和普遍必然性，自然科学也就不可能了。

②有了这些范畴未必会有自然科学，这些范畴在认识上的唯一作用就是运用于经验之上，

而经验总是包含后天偶然的成分,先天的法则和后天的经验结合起来才能构成真正的科学知识。

2) 因果性范畴在诸多范畴(“量”“质”“关系”“模态”等)中处于核心地位,具体而言:

① “量”和“质”等范畴(单一性、多数性、全体性,实在性、否定性、限制性)只是在收集和整理个别经验事实、使之构成自然科学的对象方面有用。

② “关系”和“模态”范畴则是用来规定这些已在直观中被规定好的经验对象之间的关系,以及它们与判断主体的关系,前者是对经验对象的“客观综合”,后者(可能性、现实性、必然性等)则只能是“主观综合”的。

③ “关系”在经验对象的“客观综合”中是最高的,至于就三个关系范畴即“实体性”“因果性”和“协同性”相比较而言,“实体性”是基础,“因果性”不过是实体的两种不同或相反状态相继发生的关系而已,而“协同性”(交互作用)则是因果关系的自身回转(互为因果)。因此因果性可以理解为实体范畴的真正基础,即因果关系引出了动作的概念,动作则引出了力的概念,并由此引出了实体的概念:实体并不是一个孤立的观念,而是由于它不断地向其他事物发出“动作”或“力”才被获知的。

在大数据场景下,如果说有关因果性探索的“休谟问题”给大数据分析师带来了诸如“要相关不要因果”这样高层次的思维方式指导的话,那么康德关于因果问题的探索给大数据分析师带来的则是实践层面的一些指导原则。具体而言,关于因果性的应用提示包括(但不限于):

1) 面对一大堆数据,数据分析师要首先找出不断向其他事物发出“动作”或“力”的实体,这是大数据分析的关键。

2) 找出实体或事物之间休谟所谓的“经常联结”,这在算法层次上就是类似关联规则发现的问题(对应休谟所说的对“经常联结”的一个归类),更重要的是在论域分析时,这种“经常联结”的主动发现有助于发掘大数据应用场景和技术选型。

3) 或许我们在单方面(甚至是“盲目”)地期望从数据中发现知识和洞见的时候,还应该储备和匹配的是那种康德所谓的“先天法则”,这种先天法则有的来源于应用领域本身(比如向领域专家请教和咨询),也有的来源于数据分析师自身的阅历积淀和对领域问题的敏感度和兴奋度。

4) 基于各种数据或经验发现康德所说“发生的事”的原因,或许不是大数据应用必须要做的事,但作为一种探索,这是大数据分析师应该具有的情怀,也是很多大数据项目建设中显式或隐式地进行数据后研究的哲学原因。

9.4.3 定律与模型

科学定律是用科学语言对自然界客观规律的认识进行的无歧义陈述,科学定律反映了自然界的事物、现象之间的内在的、必然的、本质的联系。但这种联系往往被各种外在的、偶

然的、非本质的现象所掩盖,需要人们运用各种科学方法将它们揭示出来。对这种联系的反映,只有经过反复验证之后,才能确立为科学定律。

科学定律具有普遍性,不同定律的普遍性程度各不相同。有的科学定律只是表述某一特殊领域内的规律,如牛顿定律(力学)、自然选择律(生物学)等;有的科学定律则表述在许多不同领域中均起作用的普遍规律,如能量守恒和转化定律等。所以,各种科学定律都有一定的适用范围,而且这种适用范围会随着科学认识的深化而发生变化。现代自然科学中的科学定律一般是用数学公式来表达的,这既方便了人们在实际中进行有效的应用,同时又便于人们运用逻辑演绎和数学演算的方法进行严格的论证。

作为一种认识成果,科学定律是客观规律的本质反映,但又只是客观规律的近似反映,是随着实践的推进及其他有关领域认识的深化而发展的相对真理。科学定律所设定的条件、适用范围、所揭示的数量关系及科学定律本身的表述形式都会在实践中不断地得到检验和修正,从而更加准确地逼近客观的自然规律。

如前所述,科学体系的构建是建立在因果律的基础之上的,因此也就使得因果性成为贯穿东西方哲学领域的一个核心话题,并引发了诸多相关哲学体系的建立和发展。本节关注的是不同领域的研究者依据不同的哲学体系框架构建的科学体系(自然科学、社会科学等)是如何作用于人们的工作和生活的,更重要的是:通过大数据建模得到的数据模型是否算得上是一个科学定律?

图9-2给出了一个科学定律在实际生活中的两个例子:

1) 第一个例子是在某个高处,手中拿着一个“石块”,一旦松手,这个“石块”就会自由落体落到地面上,这个例子中,我们看得到的“因”和“果”分别是“松手放开石块”和“落地”,而这个“因”与这个“果”之间的联系是牛顿万有引力定律(科学定律)。

2) 第二个例子是在某个高处,手中拿一个“电子”(假如我们能够做到的话),一旦松手,它的掉落路线并不能确定,甚至根本不会掉到地面上,在这个例子中,我们看得到的“因”和“果”分别是“松手放开电子”和“不能确定的掉落(运动)路线”,而这个“因”与这个“果”之间的联系是微观世界的“测不准原理”(量子物理学)。

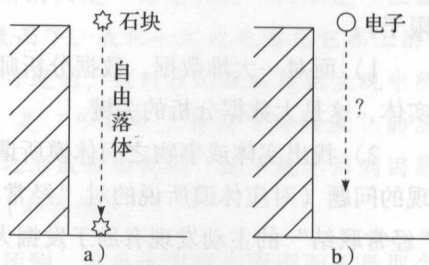


图9-2 自然规律的作用

从上面的例子中可以看出:牛顿万有引力定律反映了“石块从放手就会落地”的必然,而量子物理学的测不准原理也反映了“电子从放手就会以测不准的方式运动”的必然。因此,必然性是科学定律,尤其是(传统)物理定律的本质属性。

在大数据场景下,数据分析师可以通过诸如“松手放开石块→落地”的数据建模来得到一个预测模型,通过这个模型可以预测“放手(诸如)石块这样的东西均会落地”的可能,

值得注意的是：通过大数据建模得到的预测模型仅仅是通过对观察到的“经常联结”的“事实对”而建立起来的一个想象中的归类，反映的是人通过观察而形成的心理习惯，而不是科学定律所要求的必然性。因此传统意义上的大数据模型不属于传统意义的科学定律。

需要提出的是，大数据的工作方式恰好是：通过对历史上“经常联结”的“事实对”的建模，形成主观意义下的心理习惯，并认为这个心理习惯是可用于数据分析的（数据）后验定律，进而利用这个（后验）定律对新数据进行演绎。这种基于经验主义的工作方式和思维方式或许有别于一般意义上物理定律的应用，但这或许恰恰是不同（科学）定律应用在不同应用领域的“必然”。

佛学研究中的因果论对大数据工作者或许有一定的启发。

众所周知，佛家除了注重“因”“果”外，还讲究“缘”，其要义是：“因”能否引发“果”，需要看客观条件是否匹配，这个客观条件就是“缘”。而这个“缘”是概率性的，即有“因”未必有“果”。事实上，佛家的这种观点非常暗合康德关于“因”是“果”的必要条件的论述。

如果连接“因”和“果”之间的“缘”十分简单，简单到趋于必然（概率为1），则这种因果律就是一种决定性的因果关系，由佛学因果论引申的“宿命论”事实上就是强化了这种“概率为1”的“缘”而发展起来的；如果“缘”很复杂，则“因”与“果”的关系是概率性的，宿命论中的“改运”“补运”等做法事实上可以看作是在“缘”这个概率化的东西上下文章。

粗放地说：大数据分析师建立的大数据模型是一个输入到输出的“黑匣子”，这个“黑匣子”内部完成输入到输出的变换（通过算法），如果将“输入”看作“因”，将“输出”看作“果”，则在大数据建模的过程中尤其要注重这种“缘”：

1) “输入”到“输出”的“缘”（影响因素）有哪些？比如前文提到的“松手放开石块→落地”这个“因果”或许会受到风向、有无空气、有无其他外力等很多因素影响使得如上的“因果”在实际预测中会大打折扣，这对于应用的提示是：在大数据建模之初，在进行论域分析时，就应该尽可能多地罗列这些可能存在的影响因素，并将其作为“黑匣子”的输入纳入数据建模中。

2) “因”和“果”之间的必然（或者某个概率）是针对某一个具体的对象而说的，我们建立起的“松手放开石块→落地”这一心理习惯在用于“松手放开电子”时，就得不到“落地”的“果”，其本质原因在于建立的“松手放开石块→落地”心理习惯仅对类似“石块”这样的实体有效，这对于大数据分析师的提示在于，在进行数据建模时，必须要考虑：考察对象是谁？这些考察对象的边界是哪些？基于相似群体具有相似的购买需求的假设，我们可以为潜在的客户推荐合适的产品，这是大数据分析师应对自动推荐场景下的通常思维方式。不过这里需要注意的是：一定要界定好什么样的人群是可以归属于一类的相似人群？比如前面提到的“石头”和“电子”究竟有多大的差别呢？这就涉及前文提到的特征提取与表示问

题了,此处不再赘述,仅作应用提示。

9.5 本章小结

带来了人们极大憧憬和想象的“大数据”正在掀起一场数据技术革命,它将对人类的世界观、知识发现、思维方式及伦理道德产生全方位的影响。带着对“大数据”的好奇心,哲学家尝试从本体论、认识论、方法论、价值论和伦理论等视角对大数据进行全方位的研究,以构建一个比较完整的大数据哲学研究体系。事实上,这样的哲学体系正在形成之中。

计算机科学家从计算的角度对大数据引发的若干问题进行了科学问题的凝练,最终将问题的本质回归到计算的本质:算法+数据,并冠之以“数据科学”的称谓,还得到了工业界和学术界的普遍认可。

随着大数据在不同领域的进一步渗透及推进,在大数据部署实施过程中显现和爆发的各类技术瓶颈激励着应用研究者的研究神经,大数据技术流涉及的整个环节“数据采集→数据存储→数据分析→计算架构(模型)→系统运维”,都有不同专业领域的研究者给予关注,并因为各界的普遍投入而涌现的集体智慧使然,大数据技术持续得到推进和发展。

正所谓“应用为本”“数据为王”,大数据的价值最终体现在落地应用的部署实施上,基于思维的、基于技术的及基于数据的各家大数据公司推出的大数据产品层出不穷,各个产业角色都在基于各自不同的价值取向而支撑起整个大数据产业链。

或许,从来没有任何一个技术概念能够像“大数据”这样几乎得到“政产学研商用”各界的普遍重视;或许,也从来没有任何一个研究对象(问题)能够像“大数据”这样快速地得到从“哲学→科学→技术→应用”等不同层面的思维重建、理论重建、认知重建、价值重建。

李白在《行路难》中说:“行路难,行路难,多歧路,今安在。长风破浪会有时,直挂云帆济沧海。”

我们正处于的数据时代或许也正是如此,挑战很多、问题很多,但前途一定光明……

本章参考文献

- [1] Han J, Kamber M. Data Mining: Concepts and Techniques [J]. San Francisco, 2000, 5(4): 1-18.
- [2] Hey T. The Fourth Paradigm-Data-Intensive Scientific Discovery [M]. Springer, 2012.
- [3] Mitchell T M, Carbonell J G, Michalski R S. Machine Learning. [M]. Springer, 1986.
- [4] Russell S J, Norvig P. Artificial Intelligence: A Modern Approach [M]. Prentice Hall, 2003.
- [5] 陈波. 逻辑哲学引论 [M]. 北京: 人民出版社, 1990.
- [6] 陈飞宇. 基于集成学习算法的异常检测研究 [D]. 南京大学, 2015.
- [7] 陈晓平. 归纳逻辑及其哲学问题 [J]. 社会科学战线, 1996(4): 16-19.

- [8] 邓生庆, 任晓明. 归纳逻辑百年历程 [M]. 北京: 中央编译出版社, 2006.
- [9] 邓晓芒. 康德论因果性问题 [J]. 浙江学刊, 2003 (2): 35-42.
- [10] 贾俊平. 统计学 [M]. 北京: 清华大学出版社, 2006.
- [11] 周志华. 机器学习及其应用 [M]. 北京: 清华大学出版社, 2007.
- [12] 张志林. 因果观念与休谟问题 [M]. 北京: 中国人民大学出版社, 2010.

第三篇 大数据

实施及理性思考

• 第10章 大数据实践

• 第11章 大数据价值

• 第12章 大数据思维

第三篇 Part 3

实施及理性思考

- 第10章 大数据实施
- 第11章 大数据价值
- 第12章 大数据思维

针对个人而言，价值实现指的是外界对个体体的评价，并且这种评价往往表现为标签化

“大数据”的价值属性是社会各界对大数据的共同认知，而价值在哪里以及如何实现价值则涉及各方的价值期望及多边的交互与协作，大数据价值实现的关键在于大数据项目的成功部署实施及落地运维。本篇尝试从管理策略、价值实现及思维方式三个角度厘清大数据落地应用涉及的技术和非技术问题，并从多个视角和层面梳理各个环节的要点和细则。此外，本篇围绕“大数据实施及过程管理→大数据价值及价值评估→大数据思维及价值实现”这一主线，针对上述问题，给出各个环节的应用提示，并厘清几个基本的问题：大数据价值在哪里以及如何实现？如何部署、实施一个大数据项目？应该以怎样的思维方式和执行策略应对“大数据”的挑战？作为一个数据分析师应该具有哪些情怀？

本篇包括3章内容，分别是：

第10章 大数据实施 尝试从工程管理、技术管理、商务管理三个层面详细介绍大数据实施过程中的重要策略和思路，并从价值实现的角度依次阐述了不同视角下的进度管理问题、人员协作问题及商务管理问题。

第11章 大数据价值 尝试从大数据应用逻辑及大数据部署方式两个层面介绍大数据价值及其实现的思路和准则，并从数据本身、大数据平台两个维度阐述大数据价值的评估策略和方法。

第12章 大数据思维 尝试从数据层、分析层、应用层三个层面，依次介绍需求定位、业务梳理、建设内容聚焦、建设路径分析与规划、实施步骤及运维策略制定等大数据项目建设过程中匹配大数据价值实现及落地应用的思维方式。

【关键字】 大数据部署实施，工程管理，技术管理，商务管理，大数据价值，大数据思维

大数据实施

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的王涛、蔡洋、汤兆亮、陆恒杨、张文哲、冯艺琳等几位同学的协助，在此表示深深的谢意。

10.1 引言

《武王伐纣平话》中提及“姜尚因命守时，立钩钓渭水之鱼，不用香饵之食，离水面三尺，尚自言曰：负命者上钩来……”，说的是姜太公钓鱼的故事。

诸葛亮在《出师表》中提及“先帝不以臣卑鄙，猥自枉屈，三顾臣于草庐之中……”，说的是刘备三顾茅庐的故事。

《史记·越王勾践世家》提及“越王勾践反国，乃苦身焦思，置胆于坐，坐卧即仰胆，饮食亦尝胆也……”，说的是卧薪尝胆的故事。

本章无意点评这些经典故事本身，而是关注这些故事背后由一帮或智谋超群，或勇猛果敢，或义胆云天等诸如此类的人组成的团队，为了一个共同的价值梦想，而形成了一个价值联盟以进行价值实现，并因为后续事业的成功将这些最初的故事原型标签化为一个又一个的励志典故。其他更多的励志故事，如司马迁在《报任安书》中提及“盖文王拘而演《周易》；仲尼厄而作《春秋》；屈原放逐，乃赋《离骚》；左丘失明，厥有《国语》；孙子膑脚，《兵法》修列；不韦迁蜀，世传《吕览》；韩非囚秦，《说难》、《孤愤》。《诗》三百篇，大抵贤圣发愤之所为作也……”。事实上，我们无法建立诸如“文王拘”和“演《周易》”、“仲尼厄”和“作《春秋》”、“屈原放逐”和“赋《离骚》”之间的因果关系，但是我们可以确信的是“演《周易》”、“作《春秋》”、“赋《离骚》”这些既定事实的产生一定是有因由的，或许我们无法得悉“因”究竟是什么，但可以确信的是，这个“因”和“果”之间一定存在一个（漫长的）价值实现过程。

针对个人而言，价值实现指的是外界对人个体的评价，并且这种评价往往表现为标签化

的形式，如在组织中的领导地位、社会中的影响力等。价值实现往往是人追求的最直接的目标。

对商品而言，价值实现在于商品生产者能否成功把商品的使用价值让渡出去。价值的实现程度，或者说：这种商品能以什么价格卖出去，则取决于消费者对商品使用价值的接受程度。

在大数据场景下谈论大数据项目建设相关事宜，大数据价值应该指的是商品价值的实现，或者更确切地说：大数据项目价值实现（商品价值）的过程也是所有参与者价值实现（个人价值）的过程。

大数据是有价值的，这是所有人的共识，不过这种论调过于定性，仅仅是宏观上的一种战略意志，而能否在战术层面实现这种价值，还需要进行多方的佐证和协同，至少需要梳理和响应的问题有：

1) 大数据的价值在于面向具体应用场景的落地应用（部署、实施），但是需要注意的问题是，不是所有的应用场景都适合用大数据思维响应，关于这点应用提示在前文有所详述。

2) 大数据项目的建设需要拥有不同知识背景的角色形成联盟共同响应整个建设过程中方方面面的需求和挑战，用工程化思维管理大数据项目建设的过程是一个必然：整个建设是串行的从无到有的过程，而出于效率提升以及资源合理分配的需求，并行化也是一个必然的选择。

3) 大数据项目的建设是一个典型的技术密集型工作，因此技术管理是围绕整个大数据项目建设过程的，技术管理的目标包括：通过齐心协力，达到提高产品质量的目标；通过技术整合，达到提高团队执行力的目标；通过知识管理，达到提高团队核心竞争力的目标。

4) 大数据的价值一方面体现在产品布局的定位，这是在大数据项目（产品）建设之初就需要定位好的。在具体大数据项目的建设过程中，大数据的价值体现在目标平台的部署运行上，而所有的价值反馈均需要通过运维来实现，而运维就意味着与各个潜在协作单位的合作，商务支撑必不可少。

本章尝试从工程管理、技术管理、商务管理三个层面详细介绍大数据实施过程中的几个重要策略和思路，并从价值实现的角度顺序阐述了不同视角下的进度管理问题、人员协作问题及商务管理问题，本章的结构安排如下：10.2节从工程管理的角度介绍在大数据整个实施过程中，从思维、开发和运维三个层面需要开展的工作以及可以使用和借鉴的技术和思路；10.3节从技术管理的角度介绍在大数据整个实施过程中，从生产流程、技术流程和知识流程三个方面需要开展的工作以及可以使用和借鉴的技术和思路；10.4节从商务管理的角度介绍在大数据整个实施过程中，考虑和利用合适的商业模式的必然性以及大数据和商业模式的关系，在对一些经典商业模式进行简单介绍的基础上，给出了在开展大数据项目出的一些应用思路；10.5节对本章进行小结。

10.2 工程管理

10.2.1 思维层的应用模式梳理

大数据项目的建设者往往分为两种：一种是（垂直）应用领域的具体应用主体，另一种是完全的独立第三方，后者根据企业核心竞争力的不同，又可进一步分为数据驱动型公司、技术驱动型公司、应用驱动型公司。而事实上，数据驱动、技术驱动或者应用驱动是三种不同的思维方式，贯穿于各种类型的大数据项目建设中。我们根据大数据项目的建设主体，基于不同的思维方式，将大数据应用场景下不同角色主体的关注重点罗列如表 10-1 所示。

表 10-1 大数据应用场景下不同角色主体的关注重点

	独立第三方	应用主体单位
数据驱动	类型 A-1：通过自营平台主动生产、收集（包括采购）数据，数据是企业唯一竞争力，目标是全、广、专	类型 B-1：数据往往来源于内部自营平台，会把外部相关数据的收集纳入大数据项目建设视角
技术驱动	类型 A-2：自主研发或者在开源技术基础上进行改良，技术的泛化应用性能或者某个应用目标下的高精尖是其唯一追求目标	类型 B-2：开源技术基础上进行适应性改造、往往不首先追求技术的泛化应用性能
应用驱动	类型 A-3：寻找应用需求，往往需要对应用有足够的敏感度	类型 B-3：有明确的总体应用目标，但往往缺乏细致、无歧义的描述，实际开发中需要进行目标需求的梳理，同时目标应用场景下已经设置的物理组织结构往往会成为业务需求梳理的障碍

以下对不同角色及获益期望的大数据项目建设者的不同类别进行简单的比较分析：

1) 类型 A-1：对于以数据作为核心竞争力的独立第三方的大数据公司而言，如何收集和整理更多的数据是其唯一的获利源，因此他们在产品线的设计布局上会在以下几个方面发力（包括且不限于）：

①在数据获取方法进行技术攻关，开发有竞争力的数据采集方法。比如 ETL、爬虫工具、分布式任务调度系统，鉴于这类工具软件对其核心竞争力的重要作用，他们的此类技术往往不会公开，仅停留在公司内部使用。

②设计某种多边共赢的商务模式，借此实现多边数据的采集和整合。比如通过某种“积分机制”吸引普通人将采集的数据主动递交过来，或者通过直接购买的形式从其他数据源获得数据，或者与其他数据源拥有者达成某种协议，别人将数据递交过来，未来再从该数据的收益中获得提成。

③设计某种软件（工具），通过这些软件（工具）主动获得相关数据，比如主动构建与某个主题有关的讨论组或者群，借此捕捉加入此讨论组或者群（用户）的言论偏好，或者设计、开发某种工具（软件），用户在使用这类工具（软件）的时候主动将（用户）数据提交上来。

2) 类型 A-2: 对于以技术作为核心竞争力的独立第三方的大数据公司而言, 应着眼于研发有自主产权的、面向大数据项目建设涉及的关键技术, 并将此关键技术输送给更多的企业从而获得收益, 因此他们在产品线的设计布局上会在以下几个方面发力 (包括且不限于):

①根据公司的自主优势, 围绕大数据项目建设中的共性问题或者面向某一个领域的垂直问题涉及的难题进行技术攻关, 形成具有公司自主版权的产品。比如在 Hadoop 开源版本上进行改进, 形成性能更稳定、使用更方便的 Hadoop 版本 (面向共性问题的普适性高的产品), 或者面向社交网络自动推荐算法的设计与实现 (面向某个具体问题的产品)。

②这类公司一方面会追求算法本身的先进性, 另一方面还会追求算法封装实现的成熟性, 因此这类公司一般会鼓励员工参加各类学术性会议, 通过学术交流的形式彰显技术成果的先进性, 同时从市场层面寻求算法 (技术的呈现方式) 的典型示范应用场景, 并通过典型示范应用场景的使用效果来说明算法产品的先进性、有效性和泛化性。

③为了彰显公司的技术研发能力, 这类公司往往还会将一些主流技术以开源的方式提供出来, 借此向外界表明自己的研发实力; 或者通过一些与用户互动的活动, 将技术成果免费提供出来, 让用户在某些具体的场景下使用, 然后以竞赛的方式提高公司产品的口碑。

3) 类型 A-3: 对于以应用驱动及应用能力作为核心竞争力的独立第三方的大数据公司而言, 研发落地应用的大数据产品是其获益来源, 因此他们在产品线的设计布局上会在以下几个方面发力 (包括且不限于):

①以个人用户为研究对象, 探讨个人的需求以及个人的价值, 借此围绕 “为个人服务” 或者是 “以个人数据为基础, 为某个利益主体服务” 的指导思想开展相关数据的收集、整合、分析和应用开发, 比如通过电商数据为个人提供产品自动推荐服务或者在设计网站上为某个企业主找出对其公司产品感兴趣 (有偏好) 的个体。

②以领域为研究对象, 围绕该领域的相关数据进行整理和分析, 为该领域相关的各行各业提供服务, 比如通过实时交通数据为政府提供城市土地格局利用形势研判报告、为出行人员提供实时导航、为商家提供广告渠道。

③不论以个人还是以领域为目标对象开展的大数据项目, 一旦部署, 后期都是需要专门运维的。潜台词是说: 大数据项目的落地应用, 获益来源于运维, 因此, 大凡一个以应用驱动的大数据公司, 它必须有足够的运维能力, 才能确保大数据项目 (平台) 的稳定运行并持续收益。

4) 类型 B-1: 应用主体单位建设大数据项目, 在数据层往往具有一定的优势, 特别是在信息化建设做得比较好的情况下。事实上他们已经在数据层有了相当规模的积累, 在这样的前提下进行大数据项目的开发, 其数据层的特点有 (不限于):

①由于社会角色的分工不同, 大多数应用主体单位都会因为承担一定的社会职能而采集了一些用户行为或者行业数据等, 但一定要注意, 这些数据往往具有一定的局部性, 在大数据项目的建设过程中, 必须要考虑外部数据的补充; 事实上, 也有一些单位, 如电信、移动、联通, 实际上是储备了几乎反映用户所有行为习惯的数据 (反映社会关系的电话记录和短信记录、反映个人消费能力的消费数据、反映个人活动区域的地址数据、反映个人兴趣偏好的

上网记录等)。由于这些单位是典型的管道化企业,这意味着这些数据的使用必须得在法律和政策的框架下,依据一个合适的商业模式和应用模式来支撑。

②从外部数据源采集数据,作为本单位数据的补充,必须要考虑到数据的互补性和可连接性。前者关注的是从外单位采集的数据必须能够在数据对象的描述上有所补充;后者关注的是描述同一数据对象的数据能否建立起有效的连接,如果这点做不到,数据的互补性也无从谈起(当然这点对所有的数据采集与整合都是一样的)。

③一般建设大数据项目的应用主体单位实施大数据项目的动机往往是在本单位的数据基础之上进行更好的信息应用,从而提高本单位的服务能力,他们往往会更专注于从数据中能发现什么以及对于本单位的服务能力提升有什么益处。因此在外数据源的采集方面,往往投入不是特别巨大。这意味着在进行外部数据源的梳理和整合时,要认真评估各个数据源的数据价值,包括对本单位数据的补充能力、数据源数据的可信度等,借此保证内外数据数据的整合在一定的成本控制下能够得以实现,不过在实际操作中,这点往往是最大的困难之一。

5) 类型 B-2: 应用主体单位建设大数据项目,在数据分析层往往具有一对极端的优势和劣势,一般应用主体单位在本身应用领域具有较深入的领域知识的同时,往往对新型分析技术和手法相对陌生(信息化程度及用户参与度高的场合不是这样),这意味着,进行大数据项目的开发,在数据分析层的特点有(不限于):

①领域知识的作用至少在于两个方面:一方面,成熟的领域知识直接可以作为规则处理各类数据;另一方面,不够成熟的领域知识至少可以作为数据分析过程中康德所谓的“先天法则”,用以指导分析算法的设计与实现,同时也可以将这些领域知识作为对数据建模所获得结果的一个佐证或者用数据建模结果对这些先验经验进行补充和发展,所以对于面向领域应用的大数据项目,一个潜在的隐式需求就是为领域(专家)用户提供数据后评估支撑平台,以支撑领域(专家)用户完成自身领域知识的完善和完备。

②由于面向领域应用的大数据项目建设的动机在于更好地提高本单位的服务能力和应用水平,因此在分析手段的技术选型方面,选择最合适的技术手段即是最优。因此在进行领域应用的数据分析预研(开发)时,通常基于既有的典型技术(或者开源技术)进行二次开发和算法改良是此场合的主流,而不关注技术本身的先进性和泛化应用可能(这是完全有别于作为独立第三方的大数据公司在这个层次做的工作)。

③由于应用主体单位的主要职能是在本领域职能范围内提供相关服务,因此它的优势在于服务及服务能力方面,而数据分析及IT手段只是他们依赖的一个途径而已。但是作为领域用户,出于对分析流程和思路的监管,他们往往对数据分析的技术口径和过程有详细了解的愿望。这意味着,将数据分析手段以可视化的方式展现对于领域用户而言非常必要,这包括两个隐含需求:一是分析结果可视化,二是分析过程可视化。

6) 类型 B-3: 应用主体单位建设大数据项目,在应用层往往会陷入一个很尴尬的境地,即作为应用主体单位只有模糊的总体目标需求,而不能细致地描述具体的需求。出现这样的矛盾的原因或许在于应用主体单位往往关注的是职能定位意义下的应用逻辑,而非数据主体

意义下的业务逻辑。这意味着,进行大数据项目的开发,业务逻辑的梳理和应用模式的重组不可避免。总体而言,在应用层的特点有(不限于):

①从数据中发现潜在需求而不是用户指定需求,然后基于数据实现,是大数据项目开发和传统软件开发的根本不同。在大数据场景下,从数据中提炼出需要关注的实体对象(以及实体对象对应的属性),然后建立以这些实体对象为节点的完全图,以既有的数据为基础,探索节点与节点之间的相似性、关联性或者多个实体之间的相似性、关联性,是发掘潜在需求的基本路径。值得注意的是,在实际论域分析与建模中,我们未必将一个具体的实际物理对象建立为这样的一个实体“节点”,哪怕是一个执行过程、流程都可以建立为一个实体对象(虚拟的),然后尝试找出节点与节点之间的关系,并评估节点与节点之间的关系反映的实质,这是大数据关于相关性分析的典型思维。

②应用领域的潜在需求挖掘是一个迭代的过程,这种迭代往往是基于数据层的实体分析,但这往往会引申出对新数据源的需求(这个工作会耦合到数据采集环节)。而新数据源数据的获得,又会引起更多潜在需求的挖掘。因此,大数据项目开发中的需求变更是常态,这与传统的软件开发是完全不一样的。这也可以解释为什么大数据项目部署上线后,运维尤其重要。

③应用主体单位既有的业务部门的职能分工往往会阻碍实际应用业务的分析,或者说:既有的业务部门的业务需求往往是彼此交叉和重叠的。这意味着,在针对不同业务部门的需求调研基础上进行的应用模式的重新梳理可能会颠覆既有的部门划分。

特别需要说明的是:

1) 对于独立的第三方大数据公司而言,基于自身产品定位和优势使然,会有一个基本的思维定位,比如瞄准数据、瞄准技术研发、瞄准应用。但实际环境下,任何一个大数据公司都不会如此纯粹,它往往要兼具多方面的角色特点。

2) 作为应用主体的单位营建大数据项目的计划未必都是合理的,有许多大数据项目计划是在“大数据”概念的持续热炒过程中的盲目规划和投资。这意味着,作为应用主体的单位筹建大数据项目应该本着一个理性的态度进行(具体参见3.4节)。

10.2.2 开发层的工程实施路径

借用软件工程的定义,大数据项目的建设过程是一个典型的工程化行为。这意味着:一方面,大数据项目开发是一个集可行性分析、需求调研、系统设计与实现、运维、管理等行为的一个整体;另一方面,大数据项目的建设过程涉及各种理论、原理、方法及技术,需要用工程化的思想来指导并解决其中的各种问题,研究内容包括技术、方法、工具和管理。具体而言,大数据项目建设的主要工程实施路径包括:

(1) 项目计划

该阶段的主要任务是明确“做什么”“是否可以做”“是否值得做”并制定项目开发的计划书。具体而言:

①“做什么”指用户明确项目建设的初始目标、投资单位以及价值获益预期,其目标是

为后续所有的项目参与者提供共同的项目认知。

②“是否可以做”是从数据保障、人力保障、技术保障、物理条件约束等多个角度研判该项目是否可以。具体包括：数据可行性、经济可行性、技术可行性、社会可行性、人员可行性、法律可行性等。

③“是否值得做”是从市场占比及项目价值度的角度研判该项目执行的投资收益比是否符合建设方的价值期望。

④制定项目开发计划：这个步骤是该项目经过上述论证，被认为“可行”且“值得开发”后为本项目的未来开发制定计划表和进度预期表。值得注意的是，大数据项目往往是需要分期部署执行的，这不仅仅是因为大数据项目建设本身是一个复杂的过程，存在很多的迭代反复，除了从项目执行的可行性角度而言必须分期执行而外，还有一个重要原因是，大数据项目的开发是一个耗费人力、物力和财力的过程，必须通过尽快地部署实施让项目建设单位获得预期收益，从而为后续的滚动投资赢得充分信任；同时让项目分期部署，先部署的项目具备“造血功能”（获益），也会为后续的建设提供投资来源（主要是给投资方带来信心）。

（2）项目开发

大数据项目的开发涉及的关键环节包括：需求调研和分析、方案设计和选型、项目开发和测试、需求迭代分析和研判等。

①需求调研和分析：基于大数据项目建设的原始动机，从大数据项目涉及的人、基本业务需求、职能需求定位等角度，利用10.2.1节提及的应用模式梳理方法对潜在应用目标进行归纳和整理，同时对数据源的现状和需求（需要从外部采集或者购买）进行整理，形成基本的目标开发需求。在需求调研和分析阶段，还需要考虑的是，大数据项目建设既有的基础设施基础是什么？未来的服务平台、运维平台分别有哪些？有哪些渠道资源可供未来开发阶段使用（这些渠道包括数据获取渠道、市场运作渠道等）？彼此依赖关系如何？所有这些问题都是大数据项目开发过程中的必然限制条件。

②方案设计和选型：大数据项目的技术流程很清晰，必然是“数据采集→数据存储→数据分析→系统开发”，这个步骤需要根据上述步骤约定的目标需求以及实际的条件限制选择合适的技术手段。

③项目开发和测试：本步骤的目标是将选择的技术手段装配成满足明确目标需求的计算流，并根据对实际场景的数据测试，进行算法的改良；并将所有匹配目标需求的若干计算流整合到一个统一服务平台框架以便于用户管理和使用。

④需求迭代分析和研判：大数据项目的开发有别于传统软件开发的重要一点是，大数据项目开发的需求变更是必然的。由于大数据项目建设中的一个潜在思路是数据服务和计算服务，因此每次需求迭代引发的程序改变往往会体现在一个计算流的配置和修改上，不会对其他模块和部件产生影响，因此，从需求迭代的响应上来看是可控的。

（3）项目运维

大数据项目的建设价值获益来源就是运维。一个大数据项目部署运营后，需要跟进的就



是运维，具体而言包括以下几个层面的运维：

①软件运维：即传统意义上的软件维护，根据实际运行过程中的情况对大数据项目（本身也是一个软件产品）进行纠错性、预防性和完善性维护。

②技术运维：大数据项目是一个典型的以数据驱动为基础、以数据分析为核心的技术流，在系统运行过程中，数据的变化会使得在原先技术选型意义下进行的数据建模不再适用，必须根据实际的运行情况进行数据模型的更新，甚至重新进行技术选型。

③价值运维：这是最为关键的一个环节，大数据项目所有的获益都要通过这个环节实现，通过大数据项目平台的有效运行，获得收益，此间包括从外协单位购买数据的成本消耗以及将数据或者计算（往往以服务的方式）销售给外协单位获得的收益等。

综上所述，与软件工程相比，大数据实施在开发层次的工程管理差别如表 10-2 所示，此处不一一赘述。

表 10-2 软件工程与大数据实施

阶段	重要环节	软件工程	大数据实施
项目计划	问题定义	明确定义用户（甲方）是谁，目标是什么等	明确定义用户（甲方）是谁，应用场景是什么等
	可行性分析	从技术、人力、法律等方面论证是否可以做，以及是否值得做	除了与软件工程类似的评估指标外，还需要理性评估目标场景是否合适用大数据思维加以响应
	项目计划	按照既定的目标进行项目的分期部署，往往有明确的理论或者方法精准预估工作量、工作进度、投资回收期、产品推广方式	往往是分期部署，每期的部署都应该具备“造血”能力，项目的收益来源于系统的运维，研发的中间产品往往也可以获益
项目开发	需求分析	甲乙双方在沟通、协同的前提下完成对显式、隐式需求的无歧义表示，需求来源于用户	往往只有模糊的需求，甚至只有应用场景的描述而没有明确的需求表示。更多的情况下，需求是在开发的过程中新增和引导出来的；需求来源于数据
	概要设计	在需求明确的情况下，设计系统的总体结构，一般是以模块结构图设计作为标志，根据不同的目标需求，系统模块结构往往不一样	总体结构化一定是按照“采集→存取→建模→系统”进行，技术路线清晰。重点需要考虑数据的来源和存储的选型
	详细设计	按照概要设计的约定对每一个模块、每一个类的算法选型、输入输出格式等进行详细的设计	在需求的理解和延展的基础上，从数据服务、计算服务、应用服务、系统平台四个角度对各个环节的技术选型进行论证，往往需要进行测试评估以确定技术选型的合理性
	代码编写	按照设计的要求进行代码的编写	根据技术选型的约定进行算法的选取及编写，有大量的开源代码可复用，往往需要进行算法的改进和预研
	测试	用白盒或者黑盒等策略进行功能测试或者性能测试等，早期的需求分析和设计手册是评估基础	除了进行类似软件工程的测试外，还需要对数据建模性能进行测试和评估，以引发新一轮的技术迭代，实际应用效果是评估基础
项目运维	维护	纠错性维护、完善性维护、预防性维护等	除了软件工程意义上的（软件）运维外，还包括技术运维和价值运维

10.2.3 运维层的平台应用保障

从计算机科学与技术的角度来看,大数据就是“数据+计算”,因此,大数据项目的建设一般包括两个方面的内容:数据服务平台(建设)和计算服务平台(建设)。前者的目标是为本系统内部(或者第三方)提供透明、高并发的数据访问服务(接口);后者的目标是通过数据的有效处理,实现匹配应用场景的应用并以合适的方式提供给目标用户,或者将某种计算能力以服务的形式提供给第三方。大数据应用场景下数据层本身的分布、异构和海量性不仅给数据存储与高并发访问带来压力,还对数据计算的吞吐率有较高的要求。这也是以资源整合、按需分配为主要目标的“虚拟化技术”被工业界和学术界相关人士重视的原因。

所谓虚拟化是一种资源管理技术,是将计算机的各种实体资源,如服务器、网络、内存及存储等予以逻辑抽象和统一表示,然后将已整合的资源以一种与物理位置、物理存在、物理状态等无关的方式进行调用。虚拟化是实现物理资源复用、降低管理维护复杂度、提高设备利用率的关键,同时也是为未来自动实现资源协调和配置打下基础。由于在大规模数据中心管理和基于互联网的解决方案交付运营方面有着巨大的价值,服务器虚拟化技术受到人们的高度重视,人们普遍相信虚拟化将成为未来数据中心的重要组成部分。

对企业数据中心而言,服务器虚拟化技术对数据中心运营的价值正逐渐凸显,并具有“颠覆性”的技术前景。归纳起来,服务器虚拟化技术为企业带来的利益至少体现在以下两个方面:①通过对物理服务器和遗留存储平台的整合,提高了现有硬件和软件的利用率,避免了新一轮的采购,从而提高了投资回报率;②虚拟化能提高IT系统的灵活性。

虽然虚拟化技术可以有效地简化数据中心管理,但是仍然不能跳过企业为了使用IT系统而进行的数据中心构建、硬件采购、软件安装、系统维护等环节。一个以(虚拟)资源租赁为主要形式的服务模式又引起了人们的重视。信息技术的高速发展使得我们有可能将全世界的数据中心进行适度的集中,从而实现规模化效应,人们只需远程租用这些共享资源而不需要购置和维护。当然,更可能的方式是存在专门提供(虚拟)资源租赁服务的公司,这些公司自主建设机房,购买和维护计算机、存储设备、网络通道等,然后将这些资源进行“虚拟化”,再将这些虚拟化的资源按需分配给有需求的用户,这种租赁模式有一个比虚拟化更为人们熟知的名字:云计算。云计算采用创新的计算模式使用户能通过互联网随时获得近乎无限的计算能力和丰富多样的信息服务,并根据用户对计算和服务的使用量收费。

虚拟化本身并不是云计算,而是走向云计算的途径之一。虚拟化让数据中心的计算能力更具有伸缩性,供给也更为灵活,从而可以更好地为云计算服务。面向云计算的数据中心使用“池”的概念,每个池均可实现动态的资源调整,能够实现虚拟资源池中资源的动态调度,以达到调度过程中充分利用资源的目的。

虚拟化和云计算技术正在快速地发展,业界各大厂商纷纷制定相应的战略,新的概念、

观点和产品不断涌现。云计算的技术热点也呈现百花齐放的局面,比如基于互联网的虚拟化解决方案运行平台、基于多租户技术的业务系统在线开发/运行/运营平台、大规模云存储服务、大规模云通信服务等。云计算的出现给信息技术领域带来了新的挑战,也为信息技术产业带来了新的机遇。

在大数据项目建设中,虚拟化或者云计算(模式)都可被归为基础设施,基础设施的配套有无、优劣与否是大数据项目能否有效、稳定运营的重要因素,从软件功能的解耦以及软件能力的共享角度而言,数据服务和计算服务是大数据项目建设中的两个重要内容。

数据服务是一种软件服务,该软件服务将数据及对数据的操作以服务的形式进行封装并将其提供给消费者使用,数据服务有助于维持数据的完整性,并允许构建面向不同项目和应用的数据重用,从而促进数据可扩展价值的有效实现。

计算服务也是一种软件服务,该软件服务将计算能力进行封装并将其作为服务提供给消费者使用。计算服务有助于统一维护计算能力的版本,并通过服务选择和服务组合技术实现面向不同项目和应用的计算流重组,从而促进计算能力的可扩展价值的实现。

从大数据项目运维的角度而言,除了数据服务和计算服务而外,还有一类可称为应用服务的软件服务。该软件服务主要是面向目标应用领域不同子应用需求,利用上述的数据服务和计算服务构建面向子应用需求的计算流,并将此计算流作为响应子需求的软件服务。通过可配置的应用服务,可以有效减少因需求频繁更新而引发的额外工作量。

一个大数据项目的运维,除了有基础设施的保障以及上述的基本服务平台外,不可或缺的还包括人力资源平台、法务支撑平台、商务合作平台等,此处不再一一赘述。

10.3 技术管理

10.3.1 生产流程管理

如图 10-1 所示,一个完整的大数据生产流程依次包括数据采集(技术部门 A1)、数据预处理(技术部门 A2)、特征提取与标签化(技术部门 A3)、数据分析(技术部门 A4)、应用系统开发(技术部门 A5)、运维系统开发(技术部门 A6)。鉴于每一个环节的成果(数据或者计算)都可以通过为第三方提供服务得到价值的实现,因此,每个技术部门都对应于一个营运部门,负责对外合作,具体包括:数据营运部 I、数据营运部 II、数据营运部 III、计算服务部、应用服务部、系统运维部。

各个技术(研发)部门的职能分别如下:

1) 技术部门 A1 的主要职能是根据用户配置的数据源进行数据采集,根据数据源的不同采用 ETL、爬虫等不同策略和技术。鉴于潜在数据源往往是比较大的,因此当采用爬虫策略进行互联网数据获取的时候,往往需要考虑分布式方案,同时应该有一整套分布式调度工具供开发和利用。对于技术部门 A1 而言,数据营运部 I 是输入部门。

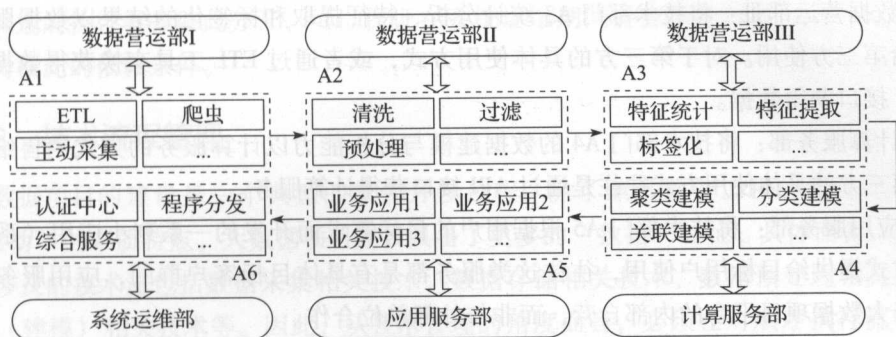


图 10-1 大数据项目生产流程图

2) 技术部门 A2 的主要职能是对采集和整合的数据进行预处理, 经过预处理的数据除了可以为本单位的后续分析及应用提供数据支撑外, 还可以通过数据营运部 II 的营运将此数据以服务的形式提供给第三方, 从而直接在数据层获益。技术部门 A2 是数据营运部 II 的输入部门以及技术支撑部门。

3) 技术部门 A3 的主要职能是特征提取与标签化, 从原始数据中提取的特征、标签和统计信息除了可以为本单位的后续分析及应用提供信息支撑外, 还可以通过数据营运部 III 的营运将此信息以服务的形式提供给第三方, 从而直接在数据层获益。技术部门 A3 是数据营运部 III 的输入部门以及技术支撑部门。

4) 技术部门 A4 的主要职能是数据分析, 可以直接为本单位的后续具体应用提供技术支撑, 事实上, 该部门的成果也可以通过计算服务部门的营运, 将此数据分析能力以计算服务的方式提供给第三方, 从而获得计算层的收益。技术部门 A4 是计算服务部的输入部门以及技术支撑部门。

5) 技术部门 A5 的主要职能是根据系统论域的分析, 综合利用上述研究成果进行目标应用系统的开发, 由于大数据项目的开展无论是需求本身还是数据本身都在不停扩张, 这意味着用户的需求会持续膨胀, 因而这个部门包括对应的应用服务部在大数据项目部署后分量更加重要, 而不仅仅是项目部署之前有价值。

6) 技术部门 A6 的主要职能是设计、实现承揽上述所有功能的统一服务平台。

各个营运 (运维) 部门的职能分别如下:

1) 数据营运部 I: 负责登记、配置潜在数据源 (URL) 以便技术部门 A1 进行数据的采集和获取, 对于需要外购的数据源, 该部门负责与合作单位进行商务合作, 并将数据交换格式配置信息提交给技术部门 A1 进行数据采集和获取; 同时该部门还负责通过主动营运 (比如与用户交互) 获得潜在数据。

2) 数据营运部 II: 将技术部门 A2 清洗、过滤及预处理后的数据通过统一整合和汇聚后, 以数据服务的方式销售给第三方使用。对于第三方的具体使用方式, 或者通过 ETL 工具交换获得数据, 或者通过 API 接口获得数据。

3) 数据营运部Ⅲ：将技术部门 A3 统计分析、特征提取和标签化的结果以数据服务的方式销售给第三方使用。对于第三方的具体使用方式，或者通过 ETL 工具交换获得数据，或者通过 API 接口获得数据。

4) 计算服务部：将技术部门 A4 的数据建模与分析能力以计算服务的方式销售给第三方使用，第三方的具体使用方式往往是通过 API 接口获得计算服务。

5) 应用服务部：将技术部门 A5 根据用户的目标需求而开发的一系列小应用（系统）以服务的方式提供给目标用户使用，往往这类服务都是有具体目标客户群的，应用服务部往往是服务于大数据项目平台的内部自营，而非与外部单位合作。

6) 系统运维部：将技术部门 A6 根据用户的目标需求而开发的统一服务平台提供给目标用户使用，往往这类服务都是有具体目标客户群的，此统一服务平台往往是服务于大数据项目平台的内部自营，而非与外部单位合作。

需要指出的是：图 10-1 是一个完整的大数据生产流程，在实际的应用场景下，由于建设甲方和获益追求的不同，实际配备的营运部门或者技术部门未必包括上述的所有环节，特此说明。

特别需要说明的是：

1) 在实际情况下，一个大数据公司往往很难是纯粹的数据公司、技术公司或者基于思维的公司，通常情况下，一个大数据公司会因为自身的优势和竞争力而向上下游辐射，从而形成更加完备和庞大的产品体系。

2) 即便以具体主体利益单位建设的大数据项目，由于项目建设本身需求，必须贯穿“数据采集→系统运维平台开发”整个流程，往往在技术积淀和数据积淀的基础上反而更加有优势，因此，很多此类公司会将其中的部分业务独立出来使其成长为一个独立的数据服务公司。

3) 理论上而言，图 10-1 中提到的数据营运部 I 从外协单位购买的数据事实上可以通过数据营运部 II 直接销售给潜在客户（合作方），但这种单纯的倒买倒卖一般不是数据公司追求的，一般数据公司根据自己的价值观通过技术部门 A2（数据预处理）实现对数据的增值（比如通过更高的数据质量实现、通过集成更多的数据源数据实现）。

4) 一般而言，数据服务公司也很难做到纯粹的数据服务，比如数据营运部 II 的合作客户从此公司购买数据后，往往需要数据营运部 II 提供针对购买数据的二次开发，这其实是数据服务型公司的产品研发提供了一个市场渠道。

5) 在一个大数据项目部署运行后，技术部门 A4 ~ A6 这三个部门往往不像传统的软件开发那样可以撤出大量的人力而仅留出部分运维人力做软件维护（往往软件维护人员是专门设置的岗位，也与开发人员无关）。大数据项目的开展麻烦在于：用户的需求本身会发生改变，或者数据驱动下的目标需求会持续滚动和膨胀，这意味着因为这种需求的膨胀而引发的所有改变均会落实到数据分析手法的改变、应用系统的增加、运维系统的维护上，因此这三个技术部门及对应的服务部门在项目部署运维后，承担的压力更大。

6) 事实上，逻辑上将大数据项目的建设解耦为数据服务和计算服务，其中一个潜在的获

利在于两边的执行事务彼此分开,可以保证在系统运维期间各方面工作的开展并行进行,而不会成为彼此的依赖条件。

10.3.2 技术流程管理

大数据项目的建设是一个典型的软件开发过程,除了兼具软件工程的特点外,由于其“数据驱动”的本质特点,大数据项目建设具备了更多的“数据”特征,具体而言,大数据项目建设涉及的技术流包括数据采集相关技术、数据存储相关技术、数据预处理相关技术、数据分析(建模)相关技术等。因此,从技术管理的角度而言,必须在对顺序执行流中的每一个环节的输入、输出做好明确的规约(因为前者往往是后者的依赖条件)的同时,还需对每个环节内在的技术内容进行有效管理。由于数据本身的动态变化,所有与数据相关的技术(每个环节内)都处于不断增加、更新和改进过程中。

(1) 数据采集方面

数据采集方面的两类主要技术是 ETL 工具的设计与实现,以及(分布式)爬虫工具的设计与实现,前者用于与自营系统或者外协单位的业务系统进行数据交换,追求的是数据交换效率更高、软件使用更便捷、软件运行更稳定;后者一般包括分布式调度系统及爬虫软件工具,分布式调度系统用于控制和调度各类爬虫工具,而爬虫软件工具则用于从互联网上获取(网页)数据。表 10-3 表示了上述几类软件在版本管理上的若干特点。

表 10-3 数据采集环节的软件工具及版本管理特点

序号	软件工具类别	功能	版本管理特点
1	ETL 工具软件	数据交换,追求稳定性、高效性和易用性	持续进行版本升级及维护,保留最新的版本(向下兼容)
2	分布式调度系统	调度控制爬虫软件工具,追求稳定性、高效性和易用性	持续进行版本升级及维护,保留最新的版本(向下兼容)
3	爬虫软件工具	从网页中获取数据,追求更广、更个性的数据获取,同时保证稳定性、高效性	1) 针对每个类型的网页或者单独个性需求的爬虫工具,需要持续进行版本升级及维护,保留最新的版本(向下兼容) 2) 上述的每一个爬虫都是彼此独立,且彼此不能替代的,这意味着每个爬虫软件都要纳入版本控制,特别注意,这个数量往往是惊人的

鉴于网页编制的手段不同(事实上是对应于网站背后的主管单位的信息化水平高低),使得无法用统一的方法(模板)从所有的网页上采集数据,同时,人们在从网页上采集数据的期望也不一样。比如,对于财务性报表,在网页中一般是以表格(矩阵)形式排版布局的,由于这些结构化信息对后续分析的重要性,因此一个潜在的要求是从网页上获取这些数据的时候,能够直接获取表格型的结构数据;再比如,对于一些政府公文性文件,虽然,其在网页上的呈现方式是一个格式规范的文本,但是其中结构化的信息其实是丢失的(比如被告、原告等),用于在此类网页数据采集的一个潜在要求是能够将整幅网页数据获取的同时,把这



些具有语义的结构化信息单独提取出来……所有这些来自数据源本身的原因或者用户对于某些类型的网页数据获取的专门要求都会使得仅仅在数据采集环节就需要维护大量的 API，而且在大多数情况下这些 API 是彼此不能替代的。

(2) 数据存储方面

在通常情况下，与数据存储相关的软件开发，大多基于既有的数据存储管理系统，对于使用 SQL 型数据库，重点在于表结构的设计，这是在设计阶段完成的，往往定下以后不再进行大的修改，一般进行配置管理，统一维护最新版本即可；对于使用 NoSQL 数据库的，则重点考虑分布式架构的设计，此处不专门讨论。

(3) 数据预处理方面

数据预处理是将原始数据进行适当的清洗、过滤以便更好地进行后续的数据建模和数据分析。一般而言，清洗、过滤都是按照既定策略利用清洗、过滤软件工具进行的。该软件工具的开发追求稳定性、高效性和易用性，因此应该对此软件工具进行持续的版本升级维护，保留最新的版本（向下兼容）的同时重点对策略的更新进行维护和跟踪（因为策略的每一轮更新都会影响后续的数据质量和目标指向的关联性）。

(4) 数据分析（建模）方面

数据分析是大数据技术流中的核心环节，其承担的职能包括特征提取（又可分为简单的统计特征提取和复杂的特征提取与选择）、数据标签化、数据建模等，表 10-4 表示了上述这几类软件（工具）在版本管理上的若干特点。

表 10-4 数据分析（建模）环节的软件工具及版本管理特点

序号	软件工具类别	功能	版本管理特点
1	统计特征提取	按照既定策略提取数据的统计特征，追求稳定性、高效性	持续进行版本升级及维护，保留最新的版本（向下兼容）
2	特征提取与选择	在数据理解的基础上进行特征提取与选择，追求的是个体多样性（往往与策略相关）、稳定性、高效性	1) 针对每一类策略驱动的特征提取和选择工具软件持续进行版本升级及维护，保留最新的版本（向下兼容） 2) 记载并跟踪不同策略更新对应的软件工具及后续的数据应用流
3	标签化	为实体对象、事件甚至数据本身从数据物理底层到高级语义层进行标签化处理，追求稳定、高效性	1) 持续进行版本升级及维护，保留最新的版本（向下兼容） 2) 标签化工作本身是与数据建模相关的，标签池需要人为配置，需要维护
4	数据建模	在特征提取或标签化的基础上进行数据建模，追求目标驱动（每一个目标对应一个建模方法）、稳定性、高效性	1) 针对每一个目标驱动的数据建模工具软件持续进行版本升级及维护，保留最新的版本（向下兼容） 2) 记载并跟踪不同目标对应的算法工具库及对应的训练集、算法模型等

在数据分析（建模）环节，由于所有的操作几乎都是与使用的策略和针对的目标相关的，

因此这个环节需要配置管理和持续维护的工具软件（无论是种类还是数量）都是最多的，此处不再一一赘述。

特别值得一提的是：标签化可以看成是在特征提取与选择基础上进行的一次数据建模，不过针对标签化的数据建模往往只会关注数据对象、事件或者数据本身的某一方面特征。比如在电信运营商手机自动推荐场景下，我们可以根据用户历史上更换手机的机型及频繁度为对应的用户打上诸如“发烧友”“商务机爱好者”“娱乐机型爱好者”之类的标签；我们也可以从用户每月的消费记录的角度为对应的用户打上诸如“高消费”“中档消费”“低消费”之类的标签。事实上，基于这些标签组合（具体实践中，远不止这些标签维度），就可以进行手机的自动推荐（推荐给有换机需求的用户）。

通过上述的简单实例可以发现，对于标签化这件事情，有的是基于简单的统计值（比如“高消费”“中档消费”“低消费”之类的标签），而有的是需要进行相对复杂的数据建模的（比如“发烧友”“商务机爱好者”“娱乐机型爱好者”之类的标签）。

除了上述涉及算法的版本维护外，进行大数据项目建设还应该储备的基本技术包括面向云环境的开发技术、部署方法、数据安全技术等。所以，大数据项目建设及项目部署运维是一个典型的技术密集型管理过程，必须对相关技术的选型及完善有完备的配置管理（管理内容包括选型依据、技术口径、改进思路等）以确保后续持续的改进和优化。

10.3.3 知识流程管理

知识管理（Knowledge Management, KM）是管理学的一个分支，由被誉为“知识管理之父”的斯威比（Karl Eric Sveivy）博士在1986年首次提出，并在随后陆续得到同行的认可、吸纳和发展。不同的研究者对知识管理的理解和定义多有不同，比如：

1) 从知识管理的目标来看，知识管理就是为企业实现显性知识和隐性知识共享，从而运用集体智慧提高应变和创新能力。

2) 从知识管理的过程来看，知识管理就是为增强企业组织的绩效而有意识进行的知识创造、获取和使用的过程。

3) 从资源管理的视角来看，知识管理就是对知识进行有意识的管理，从而创造更多的企业收益。

4) 从战略、组织视角来看，知识管理就是通过一种有意识的组织形式和策略，使得知识能够最及时地传递给最需要的人，从而帮助人们共享信息，进而将之通过不同的方式付诸实践，最终达到提高组织业绩的效果。

从管理学的视角来看，21世纪企业的成功越来越依赖于企业所拥有知识的质量，在企业组织内部构建一个量化与质化的知识系统，让企业组织中的资讯与知识，通过“获取”“存储”“分享”“转移”等过程，不断地反馈到知识系统内，永不间断地累积个人与组织的知识并形成组织智慧的循环，从而为企业创造并保持竞争优势。



大数据项目建设往往是一个周期长、成本高的开发过程，整个开发过程（含项目部署后的运维过程）是一个典型的技术密集型管理过程，且涉及多个角色的有效协作。多个角色的个体为了某一个目标而有效合作的一个基础是彼此有良好的合作沟通，这种合作沟通的本质是建立彼此认同的知识共识，以便在具体的研发过程中有共同的价值观和知识观，这意味着在整个大数据项目的建设过程中，建立全体开发人员（包括运维人员、甲方用户等）相关的知识流程管理意义重大。

知识是人们在实践中获得的被认为是正确并会（重复）用于指导未来实践的认识和经验，因此有一种观点是将“知识”分为 Know-What 型、Know-Why 型、Know-How 型、Know-Who 型四种类型。其中：

1) “Know-What 型知识”是事实知识，指的是关于事实、属性等方面的知识，比如企业有多少员工、产品用什么原料、企业主营什么业务。

2) “Know-Why 型知识”是原理知识，指的是明白企业生产的原理和规律，比如为什么选用这种而不是那种材料，以及为什么生产这种而不是那种产品。

3) “Know-How 型知识”是技能知识，指的是做某些事情的技术和能力，比如熟练工人操作机器的技能、编程熟手快速编程的能力。

4) “Know-Who 型知识”是人力知识，指的是在工作的过程中，知道如果出现异常（困难）应该请教谁的知识。

在上述的知识分类中，事实知识和原理知识有可循的规律，易于传播和共享，这两类知识被称为显式知识，一般体现在书本、资料、说明书、报告、文档中；而技能知识和人力知识往往是只可意会不可言传，因而被称为隐式知识，对于一个企业而言，物化在机器设备、软件产品等上的知识以及体现在员工头脑中的意会知识是典型的隐式知识，而且在企业知识中的比重非常大，对企业发展至关重要。在知识管理中，如何将隐式知识显式化是一个重要命题。

另一方面，大数据项目的建设过程事实上是一个内部研发团队和外部利益实体交互的过程。应当注意到，每一个环节的内外交互通常都是对目标需求、技术口径、想法、思维的交流沟通和理性趋同，这是一个典型的“知识流”创造、传播和迁移过程，往往最后的沟通结果会以显式文档、说明书等文档形式保留下来并被传播至团队（企业）内部的知识体系，而内外交互过程中所沉淀的知识（口径）是否全部被迁移至内部知识体系，事实上是无法保证的。

综上所述，从知识的来源和类型可以将知识分为四类，分别为Ⅰ型知识（内部隐式知识）、Ⅱ型知识（外部隐式知识）、Ⅲ型知识（内部显式知识）和Ⅳ型知识（外部显式知识），具体如表 10-5 所示。值得注意的是，刚才这种分类是从项目内部团队和项目外协单位的层面分为内部和外部的，而事实上，从个人知识管理的角度而言，每个“个人”可以看作内部，而其他的部分（组织级别的本单位及本单位的其他人、组织级别的外单位及外单位的合作者）都可以视为“外部”。

表 10-5 大数据项目建设的知识类型

	内部	外部
隐式知识	I 型：内部隐式知识	II 型：外部隐式知识
显式知识	III 型：内部显式知识	IV 型：外部显式知识

从资源管理的角度来看,知识管理就是对知识(资产)进行有意识的管理(具体包括隐式知识显式化、外部知识内部化、个体知识团体化、组织知识产品化等),从而创造更多的企业收益,而知识管理系统是实现知识管理的有效手段。一般的知识流程管理框架如图 10-2 所示。

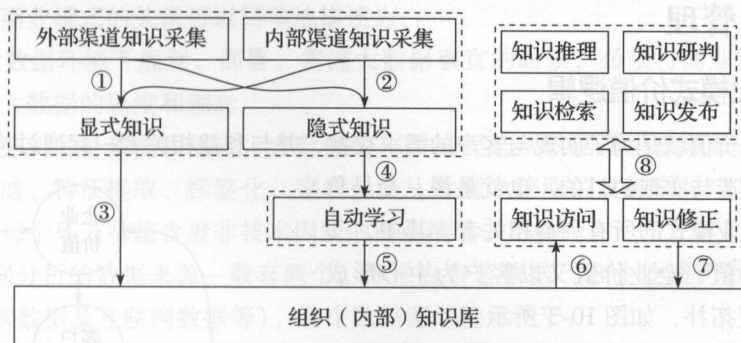


图 10-2 知识流程管理框架

参照图 10-2,大数据项目建设(运维)过程中的知识流程管理的关键流程包括:

1) 步骤①通过外部渠道收集到的知识包括显式知识和隐式知识两种,一般可以认为,显式知识可以直接存入组织(内部)知识库(步骤③)供后续使用。外部渠道的知识来源一般包括:市场人员与潜在合作方的交流、研发人员与学术同行的交流、技术人员与目标客户的交流等。

2) 步骤②通过内部渠道收集到的知识包括显式知识和隐式知识两种,一般可以认为,显式知识可以直接存入组织(内部)知识库(步骤③)供后续使用。内部渠道的知识来源一般包括:内部人员的技术交流、头脑风暴、项目复审、各类文档、员工自主学习等。

3) 步骤③是将从内部渠道或者外部渠道收集到的显式知识存储到组织(内部)知识库中供后续环节(应用系统)使用,鉴于知识本身也存在更新、修正等问题,这意味着知识库本身也需要进行维护,这点可以通过后续的步骤⑦实现。

4) 步骤④是将内部渠道或者外部渠道收集的隐式知识(不能被共享和传播)送入自动学习环节,通过学习(包括归纳、演绎等,需要专门的技术和策略,此处不再赘述)形成可共享的显式知识,然后通过步骤⑤载入组织(内部)知识库。

5) 步骤⑤将通过学习得到的显式知识存储到组织(内部)知识库中供后续环节(应用系统)使用,鉴于知识本身也存在更新、修正等问题,这意味着知识库本身也需要进行维护,这点可以通过后续的步骤⑦实现。

6) 步骤⑥为所有需要使用知识的应用场景提供统一的搜索、检索、推理等功能。

7) 步骤⑦为所有知识库更新(维护知识库)的应用场景提供统一的知识修正及更新功能。

8) 步骤⑧为所有围绕知识展开的各类应用场景提供统一的知识库存取服务。

当然,本章无意于知识管理系统的设计与实现,不过需要强调说明的两点是:

1) 在大数据项目建设及运维过程中,有必要引入知识流程管理的思路以确保大数据项目建设及运维过程中组织内部(项目团队或产品团队)的有效沟通和高效组织。

2) 知识管理系统本身就是一个大数据项目(产品),或者说,知识管理是大数据的一个潜在应用场景。

10.4 商务管理

10.4.1 商业模式价值逻辑

企业存在的价值就是为了实现与客户的需求交换,并与利益相关者一起通过构建交易关系完成客户价值的创造与实现的目的。也就是说,交易价值贯穿于整个商业模式的所有利益相关者。其中,交易价值、行业价值、企业价值又以客户为中心形成了一个商业价值拓扑,如图10-3所示。

因此,商业模式从本质上讲,是各个利益相关者之间的交易结构,此处的利益相关者包括内部利益相关者,比如某个公司实体内部的财务部门、市场部门、运维部门,也包括外部利益相关者,比如传统意义上的供应商、渠道、顾客。好的产品必须通过匹配的商业模式支撑,才有可能使得企业主体及相关利益主体最终获得收益。由于产品本身处于不断的迭代开发中,而产品所依附的环境也处在不断发展中,因此必须有动态、持续的创新商业模式才能保障各个利益主体的持续获益。正如管理学大师彼得·德鲁克所说,“当今企业之间的竞争,不是产品之间的竞争,而是商业模式之间的竞争”。

大数据的本质是在宿主环境下,在基础设施、基础平台和相关技术的支撑下,为目标应用提供解决方案(软件、服务、咨询等),从而为多边利益相关者实现各自的收益。因此大数据的价值可以从两个维度来看:

1) 从技术流来看,通过技术手段的应用实现“数据→信息→知识”的扭转,从而为具体的应用场景服务,为具体的目标应用实现“数据→利润”的变迁。

2) 另一个维度是,在整个技术流实现的过程中,由于每个环节均涉及若干利益相关者,每一个环节的利益相关者因为在相应的环节拥有相应的资源和能力而获益,如图10-4所示。

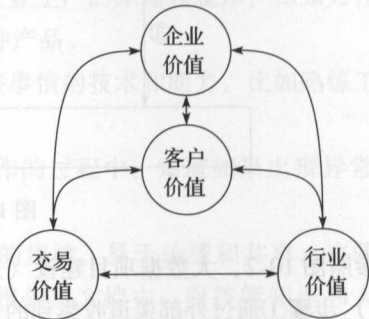


图 10-3 商业模式价值逻辑

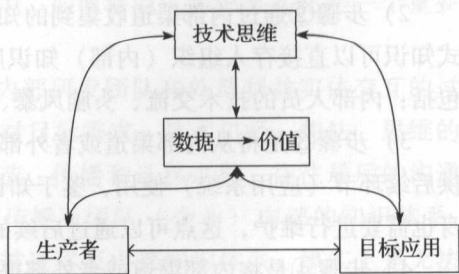


图 10-4 大数据价值逻辑

因此,大数据的落地实施不是某一方面能力和资源的独立行为,而是融合各方资源于一体的协作,最终完成“数据→价值”的实现。这意味着必须将商业模式的设计、选型、定位纳入大数据的思维和理念。值得一提的是,任何一种商业模式都不是万能的,这意味着在大数据建设及落地实施的过程中,应该根据具体的情况选择合适的商业模式,甚至创新出一种新的商业模式才能确保大数据项目的成功落地和运维。

10.4.2 大数据与商业模式

大数据与商业模式的关系可以简单地描述为:

(1) 在大数据环境下规划、部署、实施大数据事宜的时候,必须将商业模式的设计、选型、定位纳入大数据的思维和理念

以大数据分析的朴素流程而言,大数据的一般流程是从数据源中获取数据,然后对数据进行清洗、过滤、特征提取、标签化、建模,将从数据中获取的这种洞见应用在具体的应用目标中。几乎每个环节都蕴含着非技术因素的商业情节,比如:

1) 大数据分析的数据来源一般有两个,即平台自营数据(含企业遗产数据)、外部数据(含外单位他营数据及互联网数据等),各个数据源的特点在于:

①由于自营数据一般是通过自有平台收集和整理的,对于这类数据的获取或许比较简单,通过 ETL 或者 API 接口即可实现(参见 5.3 节)。

②对于企业遗产数据,由于这是企业自身的历史数据,问题也不大,只需要进行数据导入即可,虽然其中可能会涉及异构数据的处理,但总归是归属于技术口径的落实,问题往往也不会太大。

③麻烦的是在于外部数据的获取,外部数据还包括两个部分,一个是互联网数据,另一个是其他利益主体持有的数据。对于前者,在技术手段方面通过爬虫完成数据的采集是一个成熟的思路,麻烦的是其他利益主体所持有数据的数据获取,这必须通过商务手段进行交易,而如何进行交易就涉及不同利益主体之间的商务行为。最简单的是“钱-数据”交易,或者“数据-数据”交易,前者就是用钱购买数据,后者就是用自有的数据交换其他利益主体的数据,至于如何估值就是另外一回事了。

2) 看起来,“数据→信息→知识”是一个纯粹的技术流程,稍微深入地考虑一下,这个流程或许并不那么简单。出于开发效率和开发质量的考虑,目前的软件开发一般是基于构件的软件开发,其核心思想是复用。既然有可复用的构件,那么复用是最有效的一个方案。不过复用是需要成本的。可复用的构件是以软件产品的形式发布的,那么我们在开发的过程中只需购买即可(购买产品或者购买授权),这是简单的“钱-工具”交易,也是最传统和最朴素的做法,因此在很多的项目成本计算中会有一栏“第三方软件的采购成本”。而事实上,这不是唯一的构件复用方式。“计算即服务”的思想在工业界推行的时候,一种新的商业模式就出来了,即工具的使用方无须购买整个产品,而是根据使用的次数和频率来支付相应的金额,特别是在云计算大行其道的今天,这种方式渐趋流行。

3) 系统开发及运维这块更是这样,传统的系统运维一般是建立自己的机房、购买自己的服务器、购买自己的数据库等,这是一种模式。而目前更流行的模式是云服务模式,即最终用户无须购买具体的硬件资源甚至软件资源,仅需购买相应的服务即可。这对于大数据项目的提示在于:整个数据中心、运维平台均可部署在云端,数据虚拟化和计算虚拟化工作完全由云端来实现。即便对于软件开发本身,可以自己开发,可以外包开发,还可以采用“众包”(见后文具体介绍)的思路进行开发。

4) 在整个数据分析的流程方面,我们可以发现,数据的拥有者可以通过提供数据获得收益;将更多的数据收集到一起,获得更多的数据后,就可以通过提供更多的数据获得更丰厚的收益;有了数据,通过对数据加以清洗、标注之后提供更有质量的数据也可以获得收益,再或者通过对数据加以标签化,标签本身也可以获得收益;技术的提供方通过提供技术或者技术方案获益;参加“众包”的独立的人因为参与了“众包”也可以获益,当然这种获益可能是即刻的“钱-物”交换,也可能是通过未来更多利益的利益分成获得长期的收益。所有这些都意味着大数据落地涉及的每一个环节都离不开一种合适的商业模式的支撑。

(2) 商业模式的设计也需要大数据提供的支持

如前所述,商业模式的核心是客户价值、企业价值、行业价值和交易价值。那么衍生出的几个需求是:客户是谁?目标市场在哪里,有多大?客户的需求及需求方式是什么?产品的战略定位是什么?在客户眼中,产品或服务给他们带来什么样的价值?利益相关者是谁?等等。保险的说法是,大数据不能解决上述所有的问题,因为商业模式的设计毕竟需要企业领袖个人或集体的智慧。但是大数据能够为商业模式设计的具体环节提供数据支撑和辅助决策支撑。

简单梳理其中的几个应用点:

1) 战略定位。

在商业模式的设计中,战略定位关注目标(新)产品的市场定位,包括客户(群)是谁及在哪里,市场在哪里,有多大,现有产品的市场布局。针对上述问题的不同回答,企业会采取不同的市场战略,参见表10-6。

表 10-6 不同产品市场布局的战略定位

	当前产品	新产品
当前市场	市场渗透: 增加在现有市场的份额 退出: 产品退出市场 巩固和维持: 保证在现有市场的现有地位	产品开发: 新产品在当前市场销售
新市场	市场开发: 以当前产品进入新市场	多元化: 以新产品进入新市场

由此可知,战略定位取决于对当前产品及目标产品在市场上的定位,具体而言:

①任何企业的任何产品在市场上的销售情况均服从产品生命周期理论,对于产品这种有规律性的发展过程,必须要有充分认识,在进行市场需求信息调查的基础上,及时掌握产品所处市场的不同阶段,以便采取相应的对策。

产品生命周期是指新产品研制成功后,从投入市场开始到被淘汰为止的整个销售过程的全部时间。市场产品运动的发展变化轨迹可以用一条曲线来描述,这条曲线就称为产品生命周期曲线(一般是对称的S曲线),如图10-5所示。

一般来说,产品生命周期可划分为四个阶段,分别是:投入期、成长期、成熟期和衰退期。投入期的主要特征是生产成本低、投入流动资金多、广告费用大,同时产品销

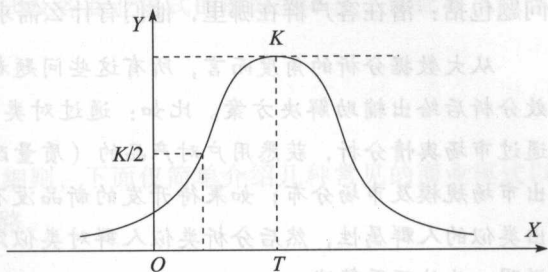


图 10-5 产品生命周期曲线

销售量增长缓慢,企业获利极少甚至为负数;产品从投入期转入成长期的标志是销售量迅速增长、利润额迅速上升,竞争者纷纷涌入,同时生产成本得到降低,生产效率和市场占有率均显著提高;成熟期是产品在市场上基本饱和,市场竞争日益激烈,销售量基本趋于稳定,利润开始减少;最后,由于成本回升、需求减少、竞争者增加和其他因素的影响,导致产品销售量减少,利润额也明显下降,产品占有率迅速降低。

这意味着产品生命周期理论是制定产品在市场上不同时期营销战略及策略的基础。在进行产品营销决策前,首先要对市场进行调查研究,做好产品定位工作,把影响产品销售的各种主要因素,纳入市场营销系统来进行分析预测。具体来说,就是认真确定企业现有业务或产品的市场现状,对每项业务和产品的战略性盈利潜力都要进行评估分析,决定哪些业务应巩固和维持、哪些应继续进行市场渗透、哪些应退出(淘汰),进而制定合理的投资计划,将有限的资金用到发展经济效益好的业务或产品中。

显然,在大数据场景下,这种辅助决策的依据就是从多个渠道采集和整合的各类数据,然后针对这些数据的分析给出一些辅助决策意见。特别指出的一点是,作为一个当事企业,不仅需要对自己的产品进行类似的分析,往往也需要对竞品进行类似的分析,借此定位企业产品在与竞品竞争的市场格局中的位置。

②任何一款产品都是针对某一个销售市场的,如果能够开发出新的销售市场(比如新的客户消费群、新的销售渠道),将企业产品投入类似的新市场中,则有可能获得较大的利益回报,这里需要解决的问题是:市场在哪里?潜在客户群在哪里?而这事实上与营销定位和商业模式定位是相耦合的。

针对既有产品开发新的市场,其关注点在于:哪些人群是潜在的客户?这些人群对销售渠道的偏好是什么?除了依赖市场人员“灵光一现”的智慧外,所有的决策都得依赖数据。从大数据分析的角度出发,一个解决方案是:在所有的既有产品销售记录中,寻找人群的聚集(从不同维度进行聚类),继而迭代分析不同维度下的人群聚集和产品的关联关系,以及此类人群的渠道偏好、消费偏好等,最终形成“在什么时间、通过什么渠道、以什么样的价格、面向什么样的消费群体开展市场活动”的辅助决策意见。

③在现有市场上开发新品是维持企业持久竞争力的重要基础,而开发新品,需要考虑的问题包括:潜在客户群在哪里,他们有什么需求,他们的渠道偏好和消费偏好如何。

从大数据分析的角度而言,所有这些问题都可以在相关数据采集与整合的基础上进行有效分析后给出辅助解决方案。比如:通过对类似竞品的分析,获悉类似产品的市场占有率;通过市场舆情分析,获悉用户对产品的(质量改进)期望;通过对类似产品的销售分析,给出市场规模及市场分布;如果待开发的新品没有类似的竞品,则从目标产品的定位出发,找出类似的人群属性,然后分析类似人群对类似定位新品的偏好。后文中有较为详细的分析和说明,此处不再赘述。

④开发新的产品投入到新开拓的市场中是企业抢占市场以便持续获利的重要举措,具体做法涉及上述第2和第3个命题,此处不再赘述。

2) 营销定位。

在商业模式的设计中,营销定位关注于客户的需求,例如:服装营销关注客户的装扮需求,饮料营销关注客户的解渴需求。针对同一需求,不同的企业又提出了不同的价值主张,在同样的价值主张下,客户又会形成自己对产品或者服务价值的自我判断(价值主张和价值感受不一样)。因此,从企业价值定位的角度出发,需要弄清楚客户的需求是什么?企业产品和服务的价值主张是什么?在客户眼中,产品和服务给他带来了什么价值?

以满足用户装扮需求的服装业为例,显然,装扮是用户购买的第一需求,在这个需求主张下,用户还有哪些细分需求?企业的价值主张是什么?针对前一个问题还涉及用户的众多细分需求,是功能性的?还是用户体验性的?如果是功能性需求,则需要继续分析类似功能的竞品是如何设计、实现的,用户的舆情体验如何?产品的质量改进意见如何?所有这些问题可以在数据采集的基础上,通过对市场舆情分析、竞品舆情分析等进行研判解决。相关的案例在13.4.4节中有较为详细的分析和说明,此处不再赘述。

本示例关注的是时尚界的Fashion流行趋势分析问题。在时尚界,为了规避失策的风险,一些零售商正在加大对潮流预测的投入,目标是提供潮流趋势的前瞻性预测,并用数据为颜色、布料、裁剪方式等方面的选择提供意见,还可以提供定制化的增值咨询服务,基于此,为时尚企业带来快速增长的同时,节约了线下调研的成本,也降低了风险。从大数据分析的角度而言,针对此应用场景和目标需求,需要线上线下采集相关数据(销售型数据、用户评论型数据、时尚媒体趋势分析型数据、一线工作人员的市场反馈等),然后通过对这些相关数据进行有效的分析,为时尚公司了解瞬息万变的时尚界的实时信息提供重要支撑。

3) 商业模式定位。

如前所述,利益相关者包括传统的消费者、供应商、服务提供商、渠道。商业模式定位关注满足利益相关者需求的方式,那么就涉及企业对利益相关者的价值主张,比如“新”、“奇”、“特”、性能、便利性、定制化、设计、身份定位、价格、成本削减、风险管理。与价值主张对应的,利益相关者也会对企业所主导的方式产生价值感受,同样是“新”、“奇”、

“特”、性能、便利性、定制化、设计、身份定位、价格、成本削减、风险管理等。

综上所述,有必要认真地梳理一下商业模式及各商业模式的脉络和本质,借此在大数据部署实施过程中加以统筹和规划。

10.4.3 典型商业模式示例

本章无意专门介绍商业模式的类别及相关细则,下面仅简单介绍几种常见的商业模式以及在大数据项目建设过程中可以借鉴的一些思路。

(1) O2O 模式

O2O(Online To Offline)这个概念最早来源于美国,是指将线下的商务机会与互联网结合,让互联网成为线下交易的前台,这样线下服务可以通过线上揽客,消费者可以通过线上筛选服务,还有成交可以在线结算,最重要的是:推广效果可查,每笔交易可跟踪。2013年O2O进入高速发展阶段,很快达到较大规模。

整体来看O2O模式运行得好,将会达成“三赢”的效果:

1) 对本地商家来说,O2O模式要求消费者线上支付,支付信息会成为商家了解消费者购物信息的渠道,方便商家对消费者购买数据的收集,进而达成精准营销的目的,从而更好地维护并拓展客户。通过线上资源增加顾客并不会给商家带来太高的成本,反而带来更多利润。O2O模式在一定程度上降低了商家对店铺地理位置的依赖,减少了租金方面的支出。

2) 对消费者而言,O2O提供丰富、全面、及时的商家折扣信息,能够快速筛选并订购适宜的商品或服务,且价格实惠。

3) 对服务提供商来说,O2O模式可带来大规模高黏度的消费者,进而能争取到更多的商家资源。掌握庞大的消费者数据资源,且本地化程度较高的垂直网站借助O2O模式,还能为商家提供其他增值服务。

有另外一种类似于O2O模式的商业模式:O2P。O2P类似于O2O,又区别于O2O。它与O2O模式的区别就是线下消费:通过网站或者移动端了解相关资讯后,再到线下的商家消费。消费者可在简单的了解之后再决定消费与否或在体验之后再支付,该类模式很适合大件商品的购买和休闲娱乐性消费,具体内容不再赘述。

在大数据场景下,针对O2O这一商业模式的应用提示至少在于:

1) 对于O2O商家而言,其价值使命在于:利用线上线下快速互动的能力,在合适的时间、将合适的商品以合适的价格、通过合适的渠道提供给合适的(潜在)消费者。从计算机的角度而言,这是一个典型的自动推荐系统,不过自动推荐系统所依赖的数据必须要实现数据采集和分析的闭环,即:①线上线下数据均收集。②线上渠道收集的数据要覆盖消费者所有的消费行为和偏好。③线下数据往往反映了用户的消费偏好、行为轨迹等,对其如何重视都不过分。④线上线下数据往往需要通过一些商业交互活动进行收集,往往通过交互活动收集的数据更具有目标指向性。⑤O2O模式采集的数据除了能够实现企业的价值目标外,或许还可以将类似的数据及分析结果提供给第三方,从而发挥数据的可扩展价值。

2) 对于大数据项目建设者而言,数据的采集与整合是其中的重要基础。前文第5章给出了相对详细的数据采集与整合方案和策略。不过从某种意义上而言,前文提及的多种数据,包括内部数据、外部数据、互联网数据等都是一种广义的线上数据,而线下数据的获得往往也尤为重要。通过与O2O平台合作获得相关线下数据或者通过自主营运交互活动主动获得线下数据这两个策略都是可选的执行思路,鉴于线下数据的重要性,在进行大数据项目建设时,应该有意识地考虑如何采集反映目标对象的线下数据,而不是仅从ETL和爬虫的角度进行数据获取。

(2) 众包模式

“众包”一词最早出现在美国,指的是一个公司或机构把过去由员工执行的工作任务,以自由自愿的形式外包给非特定的(而且通常是大型的)大众网络的模式。众包的任务通常是由个人来承担,但如果是需要多人协作完成的任务,也有可能以依靠开源的个体生产形式出现。

维基百科是一个内容自由、任何人都能参与,并有多语言语言的百科全书协作计划,其实质就是一个互联网内容领域最成熟的非商业化众包,只不过其依赖的是众多参与者的热情而非简单的金钱方面的获益(用户享受贡献内容的快乐,或者自己的贡献获得认可后的愉悦感本身也是一种收益)。

猪八戒网(www.zhubajie.com)是中国互联网众包模式的一个典型代表。该网站拟打造的价值形象是:高端Logo、VI、画册品牌设计专场,该网站一方面通过注册的方式网罗了大批各个行业专业人才,另一方面面向公众企业、机构或者个人进行宣传推广。在这个平台上,需求方发布“任务”并具体标明价格,服务方(参与众包的个体)完成任务提交方案,满意就付费,不满意就重来。

猪八戒网是一个典型的集聚众人智慧的协作平台,各方均获得相应收益。众包这种模式的典型特征是:对于发包方而言,干活的都是志愿者,所支付的佣金往往很少(一个潜台词是更侧重干活的人的兴趣指向);另一方面,对于干活的人而言,有明确的目的性,因此一般不会敷衍。这两个特点造就了“众包”的持续热度。

图像验证码是目前大型网站主流的遏制单一节点频繁访问网站以实现负载均衡的一个手段。显然这也是用户和网站交互的一个重要渠道,某些机构利用这个入口做了很有趣的“众包”项目。大致的场景描述是这样的,网站的经营方有一些陈旧且破损的资料(比如一段模糊不清的文件扫描图像),如果单纯靠公司的力量来辨识每一个文字看起来是一件既费时又费力的活动。基于众包的思路是这样:首先,将这个图像(为叙述方便,暂时仅以一张图像为例)切分成大小不一的小图像片段(确保每一个图像片段有一个有效的字符或者文字),然后为这个图像片段设置一个ID。然后将这个带有ID的图像片段以图像验证码的方式发布出去,对于访问网站的网民而言,他出于访问网站的目的应该会非常小心且认真地识别其中的字符或者文字(不认真或许也无所谓)。依据这样的做法,每个公司方就为每个ID的图像片段收

集了很多识别结果,后续利用集成的思路就可以对ID图像片段有一个精准的识别结果,这是一个典型的利用众包的思路进行劣质图像文字识别的案例。

当然这个应用场景还可以继续优化设计,比如上述的任务是利用“人们想访问这个网站”的兴趣来完成的一次协作,如果增加一些其他激励,或许这个众包的故事可以更好地滚动下去。

上述的众包思路被标签化为“众包1.0”,随着科技的发展和应用的推进,SoLoMo营销模式逐渐流行起来,在此基础上,“众包2.0”逐渐流行起来。此处的SoLoMo指的是社会化(Social,指的是以微博、微信为代表的虚拟社会交互平台)、本地化(Local,指的是基于LB的地理位置的服务)、移动化(Mobile,指的是基于以手机为代表的移动终端)的新型市场模式。“众包2.0”是指在传统众包模式上,综合应用移动Web应用、社交分析、大数据、云计算、基于位置的服务、互联网支付、O2O等技术,通过平台化管理实现双向众包或多向众包的新型商业模式。

牵牛招聘——牵牛网旗下的一款“众包2.0”概念的移动招聘软件,围绕广告传媒、游戏、IT互联网等垂直行业企业,提供“圈内人”招聘服务。其基本流程是:企业招聘悬赏众包,猎头或个人接单,自荐或内推相关人才,获得悬赏佣金,从而形成一个双向众包闭环。

当然,也有对众包模式进行质疑的声音,一个典型的观点是:众包的实质是管理的问题而非商业模式的问题。这是商业模式的研究范畴,本文不再赘述。本文关心的是:大数据在众包中能够发挥怎样的作用?

在大数据场景下,针对“众包”这一商业模式的应用提示至少在于:

1) 从众包平台的角度来看,众包的本质是提供一个协作平台使得征集者给出具体任务包时能够分而治之,本质上这是一个分布式合作求解的话题,不同的是:整个协作计算的环境是开环的,即不是网络中所有的节点都有能力、有意愿、有时间和有兴趣参与类似的协作。因此将任务泛泛地发布给全网所有节点并期待参与协作的节点能够通过自发的涌现来实现集体的智慧,即使可行,效益也应该不高,或许这也是很多理性的研究者对众包模式的担忧。但是如果我们能够基于大数据分析,对每个个体的能力、偏好、习惯、兴趣均有细致的刻画,基于此,当征集者发布任务时,平台能在合适的时间甚至合适的地点,通过合适的渠道将任务信息发布给合适的人(有能力、有信誉、有兴趣、有时间等)。这算是一种任务信息的自动推荐,对于众包而言,显然无论是效率和结果都是大有裨益的。

2) 从大数据的整个技术流上来看,很多环节都可以考虑众包这样的商业模式(思维),比如:①在数据采集阶段,将数据采集需求以任务招标的形式发布,召集更多有兴趣、有能力、有时间的人(群)进行,这是将数据采集任务分而治之的有效策略。这种做法未必适合普通的大数据项目建设场景,但应该适用于平台型数据集市进行数据采集与整合的场景。②在数据分析阶段,可以将数据分析的算法以预研任务的形式发布,召集更多有兴趣、有能力、有时间的人(群)进行,这事实上可以看成是一种预研管理平台,这种做法未必适合普

通的大数据项目建设场景，但应该适用于开源算法（项目）的设计与实现。

事实上，经典的商业模式及互联网环境下的商业模式有很多，本文仅摘取其中的 O2O 及众包模式进行分析和研判，其他内容不再一一赘述。

10.5 本章小结

从有建设大数据项目的动机伊始，大数据实施的相关工作实际上就已经展开，大数据项目部署实施的特点在于：

1) 大数据项目的部署实施从逻辑上可以划分为计划阶段、开发阶段、运维阶段，但事实上各个阶段的边界非常模糊，尤其是后两者，从大数据项目开发伊始，运维的工作就已经开始。

2) 工程管理、技术管理、商务管理是大数据项目部署实施过程中的三大体系保障，工程管理专注于大数据项目是否以及如何能够按照计划有条不紊地开展；技术管理专注于大数据项目实施团队内部如何有效地进行协作；商务管理专注于大数据项目的开发及运维过程中如何能够更好地获益。

3) 从工程管理的角度来看，大数据项目的建设始于针对应用场景的有效分析。需求分析很重要，不过应用模式的梳理更加重要，需求是数据驱动的，而非仅仅由用户驱动；这意味着在大数据项目建设过程中，需求的变更是常态，而且是必需的；在开发层将大数据项目逻辑上划分为数据服务、计算服务和应用服务，不仅是对新技术（比如服务计算）的追求，也是为了便于响应需求变化的一种策略。

4) 从技术管理的角度来看，任何一个大数据项目的开展一定是角色分工不同的一群人形成的一个组织（团队），基于对项目目标的理解、共识以及各个角色的自身特点而进行的一个长（短）期协作，协作的目标是完成大数据项目的实施，而知识流程管理是组织（团队）竞争力以及产品质量的重要保障。

5) 从商务管理的角度来看，从大数据项目启动伊始，项目运维（营运）事实上已经展开，早期或许会专注于如何以更合适的价格从更多的渠道获得更多的数据，而随着项目的推进，或许还会专注于如何以更合适的价格将大数据产品从更多的渠道辐射给更多的用户，从而使大数据项目具备“获益”能力，这是大数据项目的最终目标，也是大数据项目得到不断认可和滚动支持的重要缘由，而与外协单位的交互基础是合适的商业模式的支撑、法律法务及基础设施平台的支撑。

本章从工程管理、技术管理和商务管理三个角度介绍和分析了大数据实施过程中应该具备的一些思维方式，或许还有其他角度的一些体系保障策略，本章不再赘述。总体而言，大数据的部署实施是大数据这一个概念不断得到热炒的动力，也是大数据必然的理性回归。大数据部署实施过程中涉及不同知识背景人员的协同作业，在这个协作过程中不断澄清和细化初始模糊的目标需求，并不断坐实和落实原始的价值期望，这个过程往往是复杂的、反复的。

大数据部署实施过程往往很复杂,目标需求一般很模糊,但价值目标大都很清晰,这或许就是大数据本来的样子。

夏丏尊先生在他翻译的意大利作家亚米契斯所著的《爱的教育》的译者序中提及:“……好像掘池,有人说方形好,有人又说圆形好,朝三暮四地改个不休,而于池的所以为池的要素的水,反无人注意……”

大数据项目建设及部署实施,也应该有类似夏先生的“爱”的情怀。

本章参考文献

- [1] Armbrust M, Fox A, Griffith R, et al. A View of Cloud Computing [J]. Communications of the ACM, 2010, 53(4): 50-58.
- [2] Raymond Vernon. International Investment and International Trade in the Product Cycle [J]. Quarterly Journal of Economics, 1988, 80(2): 190-207.
- [3] Rubenstein-Montano B, Liebowitz J, Buchwalter J, et al. A Systems Thinking Framework for Knowledge Management [J]. Decision Support Systems, 2001, 31(1): 5-16.
- [4] Scheer A W, Nüttgens M. ARIS Architecture and Reference Models for Business Process Management [M]. Berlin: Springer, 2000.
- [5] 胡世良. 移动互联网商业模式创新与变革 [M]. 北京: 人民邮电出版社, 2014.
- [6] 济民, 春华. 软件工程: 原理、方法与应用 [M]. 北京: 高等教育出版社, 2009.
- [7] 江晓兴. 中国商业模式创新路线图 [M]. 北京: 中国财富出版社, 2012.
- [8] 李杰亮. 基于数据挖掘技术的移动用户手机推荐系统 [D]. 南京大学, 2015.
- [9] 林子雨. 大数据技术原理与应用 [M]. 北京: 人民邮电出版社, 2015.
- [10] 鲁松. 计算机虚拟化技术及应用 [M]. 北京: 机械工业出版社, 2008.
- [11] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战 [J]. 计算机研究与发展, 2015, 50(1): 146-169.
- [12] 魏伟, 朱武祥, 林桂平. 商业模式的经济解释 [M]. 北京: 机械工业出版社, 2015.
- [13] 张波. O2O 移动互联网时代的商业革命 [M]. 北京: 机械工业出版社, 2014.

Chapter 11 第11章

大数据价值

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的彭岳、王涛、尹康、陈嘉伟、陆恒杨、徐鸣、李永春等几位同学的协助，在此表示深深的谢意。

11.1 引言

二人同心，其利断金；同心之言，其臭如兰……（摘自《周易大传》）

岂曰无衣？与子同袍。王于兴师，修我戈矛。与子同仇！岂曰无衣？与子同泽。王于兴师，修我矛戟。与子偕作！岂曰无衣？与子同裳。王于兴师，修我甲兵。与子偕行！（摘自《诗经·秦风》）

聿求元圣，与之戮力同心，以治天下……（摘自《墨子·尚贤》）

一花独放不是春，万紫千红春满园。豆角开花藤牵藤，朋友相处心连心……（摘自《古今贤文·合作篇》）

合作是人类文明历程中的不变旋律，这不仅在于蛮荒时代唯有更多的人集聚一起同心合力才有可能打败（或者抵抗）各种野兽的袭击和围困，或者获得更多的生存资源使得人类能够延续，还在于唯有通过更务实的合作和协作，参与合作联盟的人才能够获得更多的收益，使得每个个体能够生活得更好（同时整个群体向着更好的方向发展）。理性地来看，合作的基础是拥有共同的目标并获得可接受的收益（显式或隐式）。

大数据能够引起“政产学研商用”各方的共同兴趣，其原因或许在于大数据刺激到了各方的“痛点”或者说各方都期望大数据能够有效响应和解决各自“痛点”。而大数据的自身特点也决定了没有任何一个角色可以独自地（或者说是独自很好地）应付和响应大数据的挑战，这就使得大数据涉及的众多角色唯有通过合作形成一个统一的联盟才有可能有效响应大数据引发的各类挑战并从中获益。

理性地说，联盟中每一位角色（成员）都是有独立的能力、策略和价值期望的自利实体，

因而维持联盟稳定的重要条件是每位成员（角色）能够因为合作获得彼此认可的价值收益（分配），当然这个前提是每一位角色都有区别于其他角色的能力。如何确保联盟的稳定是由市场机制引领的，本章不做介绍（这也超出了本书的关注内容）。本章关注的是不同角色（以“政产学研商用”各界为例），出于不同的获益需求对大数据在价值维度上的不同期望而采取的不同策略、战略和行动：

1) 以政界为例，在国家行政职能中，最为核心的基本职能是对内的政治职能和对外的保卫职能。从大数据的视角来看，大数据几乎涉及了政府的所有职能，因此，政府层面出台了一系列支撑大数据的政策、行动和计划，其本质原因在于：大数据被认为是国家竞争力提升、公共服务及监管能力提升、国家战略资源守护、国家数据主权博弈等的重要支撑。

2) 以业界为例，出于逐利的本能，大数据被认为是提升企业核心竞争力的重要源泉，通过利用大数据，构建新型应用模式和商业模式，获得更大收益，几乎成为所有公司关注的重心和实际“痛点”。这不是某一个公司的独立观点，而是整个产业界的普遍认同，因此，围绕大数据这一产业，已经形成一个成熟的大数据产业链。

3) 以学界为例，学术研究领域的职能和兴趣或许是通过理论和技术的研究，探究这个世界中各个领域的“是什么”、“为什么”、“怎么做”、“做什么”等哲学人文、科学技术、工程实践问题，从而推进人类文明不断地前行。大数据时代的来临不仅为商业带来了巨大的价值，对于学术领域而言也具有极大的意义，一个本质的原因在于：大数据现象及其内涵的挑战将引发学术研究者思维模式的嬗变，或许会诱发新的理论技术的革新，而这正是学术研究领域本身的兴趣和价值期望所在。

出于自身获利需求，不同的利益实体均对大数据产生了同样的、极大的兴趣，一个根本的原因在于参与的各方均认为大数据有价值，能够带来利益。大数据的价值可以体现在如下几个方面（包括但不限于）：

1) 从大数据中挖掘出隐含其中的洞见和知识，借此为不同的目标应用提供服务支撑，这是从大数据“有用”的角度表明大数据的价值，描述的是“使用价值”。事实上，除了从数据中挖掘出的洞见和知识“有用”外，原始数据或者在原始数据基础上提取出的信息同样“有用”，也是有“使用价值”的。

使用价值是指物品能够满足人们某种需要的属性。使用价值从人和自然界的关系去考察，反映着使用价值的自然属性，这种自然属性在任何历史条件下都存在，是形成交换的必要条件。使用价值用于交换，则反映了交换者主观需要上的使用价值，它包含着隐藏在其中的社会属性，它由一定的经济条件和社会条件所决定，是形成交换的充分条件。

2) 而作为一种商品，数据即资产，这意味着数据可以直接以“交换”或者“销售”的形式提供给认为数据有使用价值的第三方。在这个角度上来看，大数据是有“交换价值”的。事实上，除了数据本身有“交换价值”外，从数据中提取出的信息或者继续挖掘出的知识同样具有“交换价值”。

交换价值指的是当一种产品在进行交换时，能换取到其他产品的价值。商品和任何使用物品或产品一样，都具有使用价值。但是作为商品，具有一个特殊的属性，这就是交换。所以只有交换价值才能反映商品价值，因为它既是商品的特殊使用价值，又反映了商品的本质特征。

3) 数据“有用”或者数据“有价值”是从经济学的角度描述了数据的自然属性和社会属性。而从哲学的角度来看，价值属于关系范畴，是指客体能够满足主体需要的一种效用、效益或效应关系。这意味着，作为一种客体，数据能够满足多边的主体需要驱使而进行“交叉复用”，从而能够让其“使用价值”、“交换价值”得到持续发酵。

价值是凝结在商品中无差别的一般人类劳动，即人类脑力和体力的耗费。作为一种商品（或者产品），其使用价值会在使用过程中因为磨损、损耗而逐步降低。而数据作为一种特殊的商品，其使用价值并不会因为被不断地使用而降低，往往正相反。数据在不断的使用过程中，其使用价值将不断提高，这也是大数据的一个典型特征，即“数据交叉复用”。“数据交叉复用”体现的使用价值不仅在于匹配和响应某个主体需要，还在于能够为不同的主体提供与之相对应的使用价值。这应该是大数据的另一个典型特征，即“多靶向目标应用”。

4) 从人文角度来看，大数据的价值体现为：推进多边资源的整合、刺激集体智慧的涌现，这或许是张扬“人的社会性”和彰显“集体意志”的重要手段和契机。

“大数据”不过是人类历史进程中的一个现象或是当今时代面临的一个挑战或困难。由于这个挑战太大，需要多边资源的协同与合作才有可能得以响应和解决，本质上这是一个战术层面的技术问题，而具有自然属性（在生物进化中由人的物质组织结构、生理结构和千万年来与自然界交往的过程中形成的基本特性）的人类的社会性（对人类整体发展有利的基本性质，如利他性、协作性、依赖性以及更加高级的自觉性）是维系人类立于“万物主宰”的重要保障。通过大数据进一步强化和彰显人的社会性自然是大数据最为重要的（人文）价值体现。

本章尝试从大数据应用逻辑及大数据部署方式两个层面介绍大数据价值及其实现的思路和准则，并从数据本身、大数据平台两个维度阐述大数据价值的评估策略和方法。接下来的结构安排如下：11.2节从大数据应用逻辑的角度简单介绍从数据到知识和洞见的执行过程中，每一个环节的结果具有的潜在价值；11.3节从大数据落地的角度介绍不同的部署实施模式在彰显大数据价值上的不同体现；11.4节简单介绍大数据价值评估相关问题；11.5节对本章进行小结。

11.2 从数据到价值

已经被公认的大数据的4V特征（事实上也有不同的利益单位从自身的视角定义出若干其

他的 V 特征) 中有一个 Value 特征, 即: 大数据是有价值的, 并且在几乎所有提到 Value 特征的时候, 都提及“大数据虽然有价值, 但价值密度是稀疏的”。这种说法的动机或许是出于这样的一个价值观: 大数据的价值体现在 (海量的、多数据源的) 数据的交叉复用上, 因此细算到每一个数据源的数据, 其价值比重相对较低。

在这样的价值观认同基础上, 有一个值得考虑的问题: 如果不是从既有目标应用的角度出发去评估数据的价值, 而是从数据出发, 去找出其潜在的目标应用, 是否会更多地彰显数据的价值? 当然, 这需要智慧。目前有一个难题摆在大数据分析师面前: 有了数据, 能够做什么? 特别是已经响应既有目标应用的情况下, 如何更大力度地挖掘和发现大数据的价值。这应该是每一个大数据分析师或者产品经理需要考虑和规划的问题。

本节无意探讨如何去发掘和创新需求, 本节尝试说明的是: 在整个大数据应用逻辑中, 数据本身是有价值的, 从数据中提取出的信息是有价值的, 从信息中挖掘出的知识是有价值的, 甚至是为具体目标应用设计实现的子应用也是有价值的。换句话说, 大数据的价值体现在交叉复用, 交叉复用的对象包括大数据应用逻辑中各个环节的中间成果。

笼统地说, “数据是信息的载体, 信息是知识的载体。”这句话的潜台词是:

1) 数据是描述和表示实体对象的最原始的、未被加工的、仅具有语法意义 (比如, 按照某一种格式或者规约表示) 的原始记录, 本身除了反映实体的某个维度的属性外, 不能回答任何特定的问题。

2) 信息是通过对数据进行加工处理, 使数据之间建立相互联系, 形成具有能够回答某个特定问题的、具有语义的数据 (未必是原始数据的子集)。因此, 信息的作用就是消除数据的不确定性。

3) 知识是通过对信息进行加工处理而形成的, 反映事物内在规律、能指导未来工作和实践、具有效用意义的模式。知识又可以分为显性知识和隐性知识, 前者是已经或者可以文本化的知识, 并易于传播; 后者是指存在于个体头脑中的经验或知识, 需要进行大量的分析、总结和展现, 才能转化为显性知识。

综上所述, 可以简单地理解为: 数据是一种事实的表示; 信息是去除了不确定性, 从而变得有用的数据 (针对具体目标); 知识是在信息中提取出的一种具有指导意义的逻辑模式, 该逻辑模式可以表征诸如“是什么” (事实范畴) “为什么” (科学范畴) “如何做” (技术范畴) “谁以及何时能做” (经验范畴)。

11.2.1 数据的价值

如前所述, 数据是指描述事物的符号记录, 是构成信息和知识的原始材料, 如图形、声音、文字、数、字符和符号。因此数据的本质是记录、表示和描述实体 (对象) 的载体, 也是进一步分析、评估和研究实体 (对象) 的依据。

对于一个“个体人”而言, 人的物理组织结构、生理结构和千万年与自然界交互的过程

中逐步形成的基本特征使然，每一个“个体人”无时无刻不在感觉（视觉、听觉、嗅觉、味觉、触觉、平衡感等）和知觉（空间感、时间感、运动感等）这个世界，思考这个世界（在感觉、知觉及想象的基础上进行分析、综合、判断、推理等）、作用于这个世界（语言、行动、表情等）。可以看到，人感知（感觉和知觉）的是这个世界的数据，人的思维需要的是来自感知的数据（也包括人的生理结构决定人在大脑中存储或记忆的数据，以及历史上通过学习得到的知识等，或许也包括造物主赋予人的最原始的推理技能），人作用于这个世界又产生了大量的数据。

对于一个“社会人”而言，每个人作为在集体活动中的个体或作为社会的一员而活动时表现的特征使然，人在这个社会环境中具有不同的角色（在不同社会关系下的位置性关系）、关系（个体与个体之间的依赖、合作、竞争等）、文化（地域、制度等所引发的行为规范、价值观、习俗等）和思维（社会环境下引发的群体意识等）。可以看到，作为“社会人”，每天与这个社会的交互必然需要和产生大量的数据。

更重要的是，作为一个“地球主宰”的人而言，天赋予人的使命使得人愿意自发自觉地发掘这个世界的本源，比如：这个世界是怎么产生的？人类是怎么产生的？可惜宇宙或者地球诞生之时，还没有我们人类，当今的我们也无法回到人类诞生的那个年代，这意味着我们根本无法直接感知那个年代。我们只有利用那个年代遗留下来的（或许是造物主有意留下的）数据去窥探其中的因由。

再或者，每天我们在看报纸、上微博、用微信的过程，本质就是我们依赖这些第三方的媒体，而不是我们自有的直接感官，间接地获知我们生存的这个世界或者我们身边发生了什么。当然我们也会在微博、微信、论坛等各种媒体上留下我们个体的认知、态度以及我们感知的“身边的事”，这意味着我们在通过第三方提供给我们的数据去了解周围的环境的同时，也在不断地生产新的数据。

因此，可以毫不犹豫地说：我们需要数据、离不开数据，也在不断地产生数据，每天都生活在数据中。

因此，之所以数据是有用的、有价值的，首先体现在我们了解和认知数据所表示的实体（对象）的需求，比如我们通过某个朋友在微博上发表的言论可以获悉这位朋友的状态、心情、偏好……更重要的价值体现在，这种反映我们某个朋友的状态、心情、偏好的言论数据或许会被第三方（比如商家）采集到，更可能的是，这个第三方获取到更多人的状态、心情、偏好，加以整合和利用就可以形成一个具有群体价值意识流的汇聚，从而为更多的目标应用提供服务。

在某个电商网站上购物，出于个人的消费习惯（或者是某种获益使然），每个人会写下对这个产品的使用心得（或者商家提供的购物服务的体验）：产品的哪些功能好，哪些地方不好，消费体验哪些方面不足等。有经验的（生产此产品的）厂家会把所有购买此产品的用户写下的使用体验加以汇聚，将其中的不足和改进意见加以梳理，就形成了厂家进行质量改进的重要依据；更有经验的厂家会以同样的方法收集竞品的相关售后舆情，为本厂的产品质量

控制或者开发新品提供辅助决策意见。

值得注意的是,表示、记录和描述实体对象的维度有很多,这意味同样的实体对象,或许会有不同的数据形式(或者类型)。更重要的是,每个实体对象本身是多维的和复杂的,这意味着每一个实体对象在每一个维度上遗留下的数据或许仅仅表征了这个实体在当前维度上的状态、行为和特征,因此对一个实体对象的每一个维度的数据加以汇聚,才有可能最完备地表示和描述该实体。

以每个人常用的手机为例,通话记录可以直接反映机主的交往圈、上网记录可以直接反映机主的内容偏好、APP使用记录可以直接反映机主的使用习惯、手机与不同基站的握手信息则反映了机主的行动轨迹……仅以使用手机这一场景为例,就可以从如此多的角度对机主这个实体进行记录和表示,如果采集更多反映机主的其他数据,比如在公安信息网中留下的车辆归属信息、在银行留下的存贷款和进出账流水信息、在社保系统中遗留的社保信息、在医院信息系统中遗留的就诊和体检信息,对这些数据(和其他更多数据源的数据)加以汇聚和整合,就可以得到这个人更加完备的立体画像数据,这对社会管理、商家营销等都是大有裨益的。

综上所述,数据的价值至少体现在如下几个层次(不限于):

- 1) 描述单一实体对象的数据是有用的、有价值的,单一的数据源、反映多个实体对象的数据汇聚是有直接应用价值的。
- 2) 更多数据源的数据汇集是更加有用的、有价值的,多个数据源数据的交叉复用是数据的价值得以提升和彰显的根本。
- 3) 数据的价值应该体现在多目标应用场景的数据有效交叉复用,而不是仅仅为某个单独的应用服务。

11.2.2 信息的价值

原始数据中充满了对后续分析产生直接负面影响的噪声,这意味着必须将这些噪声有效地剔除才能够使数据得以有效地利用。噪声的产生原因有很多,具体包括(不限于):

- 1) 在数据的产生过程中,有意或无意地、人为地增加了许多的噪声。

以网上厂(商)家经常采用的软文营销为例,出于公司自身利益的追求,许多商家(往往通过“水军”)会在网上发布大量(重复或者不重复)的商业软文,借此宣传本公司的产品(甚至大量发布竞品的不利传闻)。这给后续的数据分析带来了极大的困难和挑战,因为往往有效的数据被(人为、有意地)淹没在噪声中。当然,这种噪声也是相对的,如果后续的分析目标是找出“水军”或者分析竞争对手的营销策略,这些噪声往往是必要的数据源。

以门户网站新闻频道的新闻报道为例,许多情况下,新闻都是转载而来的,因此,在进行新闻事件分析的过程中,面对大量的、同质的新闻数据,是否去冗余以及如何去冗余都是数据分析师面临的一个极具困扰的问题。

2) 原始数据在采样过程中, 由于无法分拣出有效和无效的成分, 使得采样来的数据中就包含有噪声。

以孕妇体检为例, 利用听诊器获取母体的心脏运转状况的数据势必会受到体内胎儿胎心跳动的影响。如果我们的研究对象是母体, 则胎儿的胎心跳动同样会被听诊器记录下来, 成为未来数据分析的噪声。当然, 噪声是一个相对的概念, 仍以此例, 如果研究的对象是胎儿的胎心是否正常, 母体的心脏跳动及器官蠕动等也会被听诊器记录下来成为未来数据分析的噪声。

更为常见的例子是在进行网页数据分析场景下, 利用网络爬虫可以把一个网页完整地爬取下来, 很显然, 这个数据中包含了反映这个网页内容的有效数据 (暂时称为有效兴趣区) 以及反映网页框架结构的模板信息、广告、无效链接等。显然, 当我们的研究对象是这个网页的有效兴趣区时, 其他的数据都是噪声, 反之亦然。

3) 在数据的使用过程中, 出于价值期望的不同, 许多的数据被预处理和加工, 在这个过程中无意识地增加了许多噪声。

以中药方剂数据分析为例, 众所周知, 不同年代的方剂所使用的重量量纲是不一样的, 虽然都是“两”, 但不同的年代的“两”绝对值并不一样; 同时每一味药的产地不一样, 其所使用的量也是不一样的。在这种情况下, 不同数据源的数据在集成过程中, 如果不对每一个数据源数据都使用统一的处理标准, 就会引入不确定性。如果有些数据源的数据是经过第三方加工处理过的, 这意味着数据集成的数据源被预处理加工过, 这种不确定性往往就难以避免, 势必会给后续的数据分析带来很大的问题。

因此, 针对采集的数据进行有效的预处理, 为后续应用提供高质量的输入对于整个数据分析而言是大有裨益的。需要注意的是, 许多原因导致的数据中隐藏的不确定性也会增加后续数据分析的困难。这里的不确定性是专指数据层的不确定性, 比如原始数据本来就不准确或是采用了粗粒度的数据集合, 也可能原始数据是为了满足特殊应用目的 (比如出于对隐私信息的保护, 在数据收集的过程中有意进行的特定处理) 或是经过目标导向的缺失值处理。

通常意义上, 人们真正需要的不是分散的海量数据, 而是与目标应用相关的、有语义的、精简的“厚”数据, 这些语义信息往往是基于对原始数据进行预处理后的统计及高级语义的提取得来。对数据或者对目标实体进行标签化是大数据时代最常用的策略, 标签化的本质是通过对数据在表示目标实体的语义理解的基础上, 从不同的维度对目标实体添加具有 (或者不具有) 物理语义的标签, 借助此标签, 后续的分析师无须最原始的数据, 仅仅利用这些标签就可以在标签理解的基础上进行数据的建模。从对目标实体的表示层次上来看, 标签可以分为以下两种 (不限于):

1) 低级标签: 从原始数据中直接提取出目标实体相关的统计值, 利用此统计值, 即可表示目标实体在这个维度上的数据特征, 这类标签往往从原始数据中直接获得或者通过简单的统计获得。

2) 高级标签:从原始数据中提取出与目标应用相关的高级语义标签,比如属性语义标签、情感语义标签、风险语义标签。这类标签的提取往往需要借助数据建模方法获得。根据标签的适用面,可以将此类标签分为通用标签和专用标签,前者在应用面上具有相对的普适性,有利于数据交叉复用场景对标签的复用,后者是专注于应用目标而进行的有针对性的标签(或许对其他复用场景有用)。

以基于互联网金融情报采集的企业征信应用场景为例,利用爬虫从互联网(配置的数据源)获取到大量的数据以后,可以根据配置的关键词(热词),为每一个网页标注这个网页的若干标签,比如:公司名、热词、敏感词。基于这些标签化的网页,就可以在不触及网页原始数据的基础上,在标签词意义上进行关联分析。当然也可以在此基础上,为每一个配置的公司标注若干统计标签,比如这个公司在数据库中出现多少次、有多少次与哪些热词或者哪些敏感词相关,所有诸如此类的标签可以理解为是一种面向网页或者公司的低级标签。这样的标签化方法可以更深入一些,比如通过对网页的高级分析,可以获悉某个网页与哪些风险规则相关,借此为这个网页标注上对应的风险标签;也可以通过情感分析获得某个网页反映的情感倾向,借此为这个网页标注上对应的情感标签。所有诸如此类的标签可以理解为是一种面向网页(或者进一步以公司为中心进行标签化)的高级语义标签。

综上所述,信息的价值至少体现在以下几个层次(不限于):

1) 通过对原始数据的降噪、去冗余等预处理操作,使得原始数据在目标表示方面的功能更加纯粹,这有助于增加数据关联的可信性和完备性。

2) 通过从原始数据中梳理出数据在表示目标对象方面的相关性,从而降低数据在表示目标对象上的不确定性,这有助于增加分析师在数据语义表示上的明确性和一致性。

3) 从原始数据中提取出数据在表示目标对象方面的语义特征,比如统计特征、低级语义特征、高级语义特征,这些从数据中凝练出的精简特征本身就可以描述目标对象,并可以作为后续数据分析的输入,显然是极有价值的。

11.2.3 知识的价值

从大数据中发现知识和洞见,并且用这种知识和洞见指导未来的工作和实践,自然是有价值的,而这也是各界对大数据抱有极大期望和热情的重要原因。比如商家可以通过对用户的立体画像,根据历史数据建立诸如“什么样的用户会购买什么样的产品”或者“购买了某产品的用户会购买其他什么样产品”模型,借此可以为商家的精准营销和产品推荐提供支撑。

2015年8月4日,美国市场观察网站报道,美国食品和药品管理局(FDA)批准了第一种3D打印药物(药物名为Spritam,用于成人和儿童癫痫症患者的口服类药物),网站在报道这个新闻的时候关注的是“3D技术在制药领域的潜能”。这个技术未来应用的一个潜在场景是:未来药店可以根据你的身体特征指标,如体重或身体脂肪含量,打印出相应剂量的药物。问题是,什么样的体重以及什么样的脂肪含量(或者还有其他的一些属性)应该使用多少的

剂量呢?这个需要从历史数据(个体属性、使用量和疗效)中建模得出。事实上,在互联网时代常被提及的C2B制造模式,就是根据用户的需求进行产品的设计和制造,而这需要的就是什么样的人与什么样(规格)的产品对应的关系,发现这样的规律是确保C2B成功实施的关键,而这都需要从数据中发现含有如上所述的映射关系的知识和洞见。

在历史数据中通过建模寻求此类“属性数据—目标(值)”的关系并直接应用于具体的应用实践,显然能够直接为商家带来商业价值,这算是“数据→价值”的直接体现,或许更重要的价值在于:领域专家通过对历史数据及模型进行后评估研究,从中发现“数据与数据”、“数据与目标”之间关联度的“因果规律”,进而拓展和丰富领域知识,大数据在此方面彰显的价值往往更大。

在面向“数据→价值”的数据分析中,一直存在两种(往往并行使用)策略:基于领域专家的知识(经验)以及根据历史数据进行建模(机器学习及数据挖掘)。某种意义上而言,纯粹的机器学习是基于历史数据的建模,尝试得出“数据→目标”的模型,潜在的几个问题是:①这个模型的输入(自变量)是否足够完备?②模型本身是否最优?③模型训练成本(时间、空间等)是否足够匹配需求和满足实际实施条件的约束?相比较而言,基于领域知识的数据分析的优势在于:如果有完备的领域知识并且能很好地形式化到分析系统中,在效率和精度上往往会有较满意的结果。不过存在的困难和挑战在于是否具有这样的领域知识并能够很好地形式化?这个问题往往是领域专家的兴趣所在,对一般意义上的数据分析师而言,将所有数据、信息、模型以素材输入的方式提交给领域专家,领域专家依赖自身领域的特点,综合利用专门的相关性分析手法,研究出或者发掘出可形式化的领域知识,从而对领域的理解有用,对传统意义上围绕“数据→目标”的分析有用。因此一个大数据项目的开展,往往会并行一个后评估研究,专门针对领域数据展开,目标是发现领域知识。

综上所述,知识的价值至少体现在以下几个层次(不限于):

- 1) 通过机器学习和数据挖掘对历史数据建模,借助此模型完成目标应用相关的价值实现,这是从数据中发现知识和洞见的最基本应用。
- 2) 将数据模型以服务的形式提交给第三方,第三方综合利用相关模型,为更多的目标(创新)应用提供支撑,这是从数据中发现知识和洞见的应用拓展。
- 3) 通过大数据分析,将数据、信息以及模型作为研究素材提交给领域专家,领域专家结合其他手段丰富和拓展领域知识,这种知识可以为数据分析本身服务,也可以为更加泛化的领域应用服务,这是从数据中发现知识和洞见的高级应用。

11.2.4 应用提示

大数据涉及的相关利益实体,除了最终用户(来自政界、业界、学界等的自然人、法人等)这个角色外,还包括基于数据的公司、基于技术的公司、基于思维的公司这三种角色(或许有更多的角色)。基于数据的公司专指生产数据,或者通过购买、交换等采集手段而拥

有数据的公司；基于技术的公司专指为大数据实施提供基础平台保障、技术架构保障的拥有自主技术的公司；基于思维的公司往往专指对应用领域需求、技术响应具有极高敏感度，能够快速、高效嫁接需求与解决方案的公司。显然，这种划分仅仅是角色层次的划分，事实上，许多公司承担并兼任着多个角色，以谷歌为例，它既是数据公司，又是有自主知识产权并且有高质量研究团队的技术公司，同时它也有面向具体应用领域的解决方案。

在中国，产业结构调整与优化的持续政策推进以及2015年发布的“中国制造2025”计划都在倒逼中国的企业，尤其是制造型企业适应时代的趋势及国家的发展战略进行技术融合（创新）、产品融合（创新）、业务融合（创新）、产业衍生（创新），而大数据被公认为是其中一个重中之重的基础部分。

从2002年我国就开始进行的“两化融合”战略以及2015年推行的“中国制造2025”其本质都是在进行产业结构调整与优化的持续政策推进。2002年十六大报告中提出“……以信息化带动工业化、以工业化促进信息化”；2007年十七大报告中提出“……促进信息化与工业化融合，走新型工业化道路”；2012年十八大报告中提出“……促进信息化与工业化深度融合……”；2014年政府工作报告中提及“……促进信息化与工业化深度融合，推动企业加快技术创造……设立新兴产业创业创新平台，在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展”；2015年政府工作报告提出“……制定‘互联网+’行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展”；2015年5月8日，李克强总理签署《中国制造2025》（关于《中国制造2025》的详细介绍参见后文13.4节，此处不赘述）。

这意味着，在中国国情下开展大数据项目，不可忽视的一个重要参与方是中国的传统企业，尤其是中国制造企业。他们的参与方式往往不再是传统意义上的纯粹甲方，而是直接参与部署和实施的研发力量。当然这对于其他角色的利益实体的一个提示是：将大数据（技术）应用于两化融合或者工业互联网，应该是大有前景的事情。

总体而言，“从数据到价值”的应用提示至少有（不限于）：

1) “应用为本”是大数据时代的一个典型特征，离开了具体的应用场景，大数据难以从媒体热炒的“神坛”落地。值得注意的是：这里的“应用场景”泛指与大数据相关的各个领域、行业及大数据产业链涉及各个环节的不同角色的目标需求，对于最终需求方而言，不同的价值期望和角色使得最终需求方对数据、信息和知识具有不同程度和层次的需求。

2) “数据为王”是大数据时代的一个典型特征，数据的重要性不仅在于从数据中可以发现有助于实践的知识、洞见，还在于数据本身对于很多的应用场景而言已经可以直接利用。这意味着在大数据项目实施过程中，应该有一种意识，即将数据以某种方式（一般是以服务的方式）免费地或者有偿地提供给潜在受众借此获益。这涉及数据采集与存取问题，对技术侧的挑战是：用何种策略、何种方法实现数据组织的优化（索引优化以及与应用耦合的SQL优化和表结构重整等）及高效数据存取，借此为数据的需求方（含后续的数据分析）提供高

并发、透明的统一数据平台，以充分响应数据本身的海量性以及数据并发访问的海量吞吐率需求。

3) “数据为王”这一特征在强调数据重要性的同时，另一个潜台词是，原始数据中往往充满着不确定性（这种不确定性恰好反映了数据潜在的泛化应用价值）。因此，去除原始数据中（面向目标应用）的不确定性，从原始数据中提取出更“纯”、更“厚”的信息，对于不同目标应用的服务者而言大有裨益，前者关注于数据的质量是否有保障，后者关注于数据的语义是否对目标应用有意义，所有这些都需要对目标应用有充分的论域研究和分析研判。作为对原始数据的一种（抽象）虚拟化，数据标签（语义）化或许是原始数据与目标应用之间沟通的一个有效桥梁，数据标签化一方面为后续研究和应用屏蔽了数据交叉复用的复杂性，另一方面为后续的分析或研究提供了面向目标应用的语义特征。数据标签化涉及数据理解和论域研究，对于技术侧的需求是，综合利用领域知识和行业经验，从原始数据中提取目标驱动的高质量特征向量，并在此基础上提取面向目标应用的通用标签、语义标签和专有标签，借此实现对数据理解的完备性。

4) “建模为核”是大数据时代的一个典型特征，从数据中发现可以用于指导生产实践的知识和洞见是人类文明发展过程中亘古不变的行为。大数据时代的一个核心手段是数据建模，数据建模的目的除了为具体的应用目标服务外，也为领域专家的领域研究提供技术支撑，借此发现隐藏在数据中反映目标应用领域的规律、公理（或者定理）。值得注意的是：大数据项目的开展涉及多个角色的协同与合作。这意味着在大数据项目实施过程中，应该有一种意识，将数据分析手段或建模方法以某种方式（一般是以服务的方式）免费地或者有偿地提供给第三方借此获益，这涉及计算服务问题，对技术侧的需求是：用何种方法、何种策略实现多层次、多维度的计算服务及服务组合，借此为更泛化的目标应用提供可扩展的、高效的计算服务。

5) “共享合作”是大数据时代的一个典型特征，从数据中发现知识和洞见涉及不同利益角色的分工和合作。因此，大数据最重要的价值体现在于：在多方资源融合的基础上刺激集体智慧的涌现。作为一种部署实施方式，云计算将计算资源本身和泛化多方资源进行了有效的虚拟化，从而为多方角色提供了多方个性所需的、弹性可扩展的服务。因此，服务是大数据时代的主旋律，各个利益角色将各自的能力以服务的形式封装输入不同的多方实体并从中获益。前面提到的数据服务、计算服务也正是基于这种精神。

11.3 从闭环到开环

11.3.1 垂直应用价值

一个大数据项目的开展，大都是围绕一个具体的应用场景展开的。这样的思维方式基于一个朴素的价值观：基于用户需求被动地研发直接满足用户需求的（大数据）产品，进而和

用户一手交钱一手交货，大数据实施方通过一件件卖产品实现利润的加法式增长。基于这样的认识，所有的大数据项目都是从（更多、更广的）数据采集开始，然后开发面向目标应用的分析工具集，为具体的垂直目标应用提供技术支撑（往往通过一个统一的云平台）。

基于这样的价值观，大数据项目的开展一般思路有（不限于）：

（1）应用层：是否还有更多的需求没有挖掘？

在这个价值观的指导下，需要领域专家介入，与大数据项目实施方共同探讨是否有潜在的应用需求或者既有的应用是否可以重整（组）。鉴于“应用为本”这一特征，对应用需求的敏感度和业务重组能力往往决定了大数据项目开展的方向，并直接决定了大数据项目开展的规划。

（2）数据层：是否有更多的数据源？

在这个价值观的指导下，需要领域专家介入，与大数据项目实施方共同配置潜在的数据源，然后设法通过技术手段或者商务手段获得已配置数据源的数据。鉴于“数据为王”这一特征，数据采集的广度和深度往往决定了大数据项目的成败，而这通常受制于大数据项目建设方（包括甲方和乙方）的资源边界有多广和多深。

（3）算法层：是否有开源的或者更高明的算法支撑？

在这个价值观的指导下，往往需要项目实施方有自主的算法研究能力，在领域专家和用户的辅助下完成面向目标应用各类算法的设计与实现。鉴于“建模为核”（其实除了建模外还涉及其他各个环节的关键技术需要攻关和复用）这一特征，实施方的技术研发能力和选型能力往往决定了大数据项目的成败。

“应用为本，数据为王，建模为核”是大数据项目开展思路的关键所在，也是其产业发展的门槛与瓶颈。上述的大大数据项目建设思路是典型的业务驱动型的闭环开发思路，典型的局限在于（不限于）：

1) 在一个封闭的垂直应用需求场景下的数据汇聚和复用或许已经很好，但是不同利益主体建设的大数据项目又成为彼此独立的“大数据孤岛”（每个大数据系统又成为一个孤岛，类似于传统意义上一再提及的诟病：“数据孤岛”或“信息孤岛”），这不仅导致了各个“大数据孤岛”中数据汇聚的重复建设，也难以形成更“全”意义上的数据集聚效应。这直接导致数据价值评估方面缺少战略级的支撑，直接影响不同利益实体在数据权益上的获益分配。

2) 在一个封闭的垂直应用需求场景下的计算服务或许已经很好，但是不同利益主体掌握的高级分析技术多以专利或自主知识产权的形式封闭在各个利益实体，难以普适复用和共享开放，针对大数据引发的挑战，在理论技术层面还欠缺战术级的支撑，大数据价值难以彰显和体现。

3) 多边协同缺乏实质推进。大数据项目的落地实施本应该是在多边资源融合的基础上刺激集体智慧的涌现，但由于不同领域、行业的差异性太大，缺少激励相容的协作和协同机制。

11.3.2 平台集成价值

基于产品思维的大数据开发（实施）方围绕具体的（垂直）目标应用提供服务并通过为

更多的、不同的（垂直）目标应用提供（往往同质）服务获得利润的加法式增长，对于应用型开发公司而言，这或许已经是一个稳妥的发展规划。应当注意到，作为一个基于产品思维的应用型开发公司，其为每个甲方部署实施的大数据项目并非都是“从无到有”的开发，一个重要的原因是，大数据项目实施方在不同的大数据项目开展过程中可以复用各个项目开展中的成果和经验，从闭环地为某一个具体的目标服务演变为开环地为不同的目标服务，这或许才是应用型开发公司获得更大收益的源泉。

从闭环走向开环，这是大数据项目的开展和实施在完成一个闭环的任务以后应该和必须要考虑的规划。基于这样的价值观，大数据项目拓展和推进的一般思路有（不限于）：

（1）已经采集的数据如何向第三方开放以及是否能从第三方获得更多数据

在这个价值观的指导下，在进行一个面向某（垂直）目标应用开展大数据项目的时候需要考虑的问题有（不限于）：①需要哪些数据？这仍然是与目标应用相关的论域研究问题，是所有大数据项目建设不可避免的首要问题。②这些数据是否可以从第三方获得？这需要对整个第三方的数据采集市场有全面的了解和足够的敏感度。③这些数据是从第三方获取还是自行获取？这需要从项目建设成本与进度以及大数据项目的建设目标等不同维度，在大数据项目开展之初进行顶层设计。④自行获取的数据是否可以向第三方开放？这需要从大数据项目的建设背景、目标以及大数据部署运维等不同维度，在大数据项目开展之初进行顶层设计。⑤自行获取的数据如何向第三方开放？这涉及技术层如何提供第三方接入的标准和接口以及商务层如何运维。

（2）已经采集的数据还能为何些应用服务以及这些应用还需要哪些（更多的）数据

在这个价值观的指导下，在进行面向某（垂直）目标应用开展一个大数据项目的时候需要考虑的问题有（不限于）：①已经采集的数据反映了哪些物理实体？哪些应用场景需要这些问题实体的数据？这是典型的数据交叉复用的思维，需要对应用和数据储备有足够的敏感度。②对于具体的应用需要哪些数据？这仍然是与目标应用相关的论域研究问题，是所有大数据项目建设不可避免的首要问题。

（3）已经储备的高明技术如何向第三方开放以及是否能从第三方获得计算服务

在这个价值观的指导下，在进行面向某（垂直）目标应用开展一个大数据项目的时候需要考虑的问题有（不限于）：①需要哪些算法（包括框架、架构等）？这是与目标应用相关的论域研究问题，是所有大数据项目建设不可避免的重要问题。②这些算法是否可以从第三方获得？这需要对整个技术进展和第三方技术服务市场有全面的了解和足够的敏感度。③这些算法是从第三方购买还是自行研发？这需要从项目建设成本与进度以及大数据项目的建设目标等不同维度，在大数据项目开展之初进行顶层设计。④自行研发的算法是否可以向第三方开放？首先要考虑这些算法是否具有泛化性和普适性，同时需要从大数据项目的建设背景和目标以及大数据部署运维等不同维度，在大数据项目开展之初进行顶层设计。⑤自行研发的算法如何向第三方开放？这涉及技术层如何提供第三方接入的标准和接口以及商务层如何运维。

上述的大数据项目建设思路是典型的平台驱动型的开环开发思路，其关注的是如何将

(更多的)数据和技术以及(更多的)目标 and 需求汇聚在一个“池”中,然后不同的利益主体在一种激励兼容的合作机制指导下,实现不同目标需求在一个平台上的对接。

11.3.3 生态协同价值

基于平台思维的大数据项目开发的理念是:大数据涉及的所有角色在一个统一的平台上各尽所能、各取所需,一起创造价值、享受获益,也就是说,平台思维关注的是集成和融合,而集成和融合在一起的各个角色提供的服务是否能够获得各方利益的最大化,平台思维或许并不关心。其中隐含的几个值得思考的问题有(不限于):

1) 大数据的价值体现在为不同的、彼此独立或是唇齿相依的目标应用(服务对象)提供服务。在有限资源情况下,为哪一类服务对象提供服务更能彰显大数据的价值?

2) 为彼此独立或唇齿相依的目标应用(服务对象)提供服务的大数据本身涉及大数据产业链中的众多环节和角色,怎样的大数据产业结构才能最大限度发挥大数据的价值?

这两个问题涉及大数据开发过程中的另外一个思维方式,即生态思维。生态思维的价值观是:大数据可以为不同的目标应用提供服务获得价值,而如果能够为整个上下游产业链(涉及不同环节的需求)提供服务,则获取的价值更大。这样的价值观其实也很朴素:基于产品思维的大数据项目往往关注某一个具体的垂直应用场景,而如果将这个垂直应用场景放大到整个产业链,则产品思维就是生态思维。因此生态思维是包含产品思维的,不同点在于产品思维关注于围绕某个目标应用的数据交叉复用,而生态思维还会在数据交叉复用的基础上关注不同目标应用之间的相关性和各个目标应用之间的交叉复用。这对于大数据价值的彰显是大有裨益的,比如:

1) 基于生态思维的大数据项目会关注整个产业链涉及的相关数据的采集与汇聚,因此所收集的数据在(产业链)目标应用的指向性更高,数据交叉复用的价值体现会更明显;相反,如果大数据项目服务的是彼此独立的垂直目标应用,每个垂直目标应用所采集和汇聚的数据未必具有很大的彼此相关性,数据交叉复用的价值彰显度就会低很多。

2) 基于生态思维的大数据项目为整个产业链不同环节提供的服务本身往往具有很大的相关性和可复用性;相反,如果大数据项目服务的是彼此独立的垂直目标应用,每个垂直目标应用所沉淀的高明算法和应用系统未必具有很大的泛化应用可能,这意味着需要为不同的目标应用配备不同的人、财、物,使得开发成本增加,从而降低整个项目实施的获益。

3) 或许一个更重要的优势在于:通过为一个产业链提供服务而不是为彼此独立的垂直目标应用提供服务,可以重整大数据产业链本身,衍生出面向不同产业链的大数据产业角色,这对于大数据产业链的结构优化本身大有裨益。

11.3.4 应用提示

闭环和开环是大数据落地应用的两种不同情怀,本身谈不上孰优孰劣,一切均由大数据项目建设的动机决定,这与大数据项目建设方的价值期望相关。

一般而言,大数据项目建设大多是从朴素地响应垂直目标应用需求开始的,只不过在持续开发和运维的过程中,不断地衍生成为平台型、生态型项目。因此,产品思维、平台思维、生态思维可以理解为大数据项目建设的不同阶段(发展路径),谈不上孰优孰劣,只是越往后发展,其价值期望更趋向提高生产力,当然这需要长远的规划和执行力驱动,这也是许多航母公司之所以成为航母公司的原因。

总体而言,大数据项目的开展都是基于某种价值期望,围绕既定的价值目标,联合包括应用模式(梳理)、商业模式(设计)、法律法务(保障)、技术思维(支撑)等多边资源进行的一种协作和合作。基于产品思维的大数据项目朴素地专注于垂直目标应用需求响应,进行有针对性的数据采集、数据存取、数据分析(建模)、系统实现(运维);基于平台思维的大数据项目专注于各边需求、资源、服务的汇聚,通过集成实现多边价值的共赢;基于生态思维的大数据项目专注于目标应用(涉及的产业链)及大数据产业链本身的结构优化和重整,期望获得结构优化基础上的生产力提高和价值发酵。在大数据项目开始及整个推进过程中的若干应用提示(问题清单)有(不限于):

1) 做什么:价值期望怎样?用户是谁?目标需求是否明确?业务梳理是否明晰?目标需求描述的应用场景所涉及的上下游产业链涉及的业务及需求有哪些?

2) 有什么:数据源在哪里?数据在哪里?数据质量如何?数据的广度、活度、混杂度、厚度如何?技术储备是否足够?是否具备更多的第三方资源?

3) 如何做:建设内容、建设路径、实施步骤是否明确?谁以及如何运维?各方保障是否具备?

11.4 大数据评估

11.4.1 数据价值评估

“数据交叉复用”是大数据时代的一个典型特征,也是数据价值得以彰显的重要思维基础,这意味着,某个数据源的数据或许能够直接服务于某一个目标应用从而彰显这个数据源的数据价值,而这或许只是其数据价值的“冰山一角”。

传统上的思路和做法是:采集的数据一旦满足某种既定需求,我们就(可以)认为该数据的采集和分析已经达到了初始的目的,这些数据就可以删除了。这种思路和做法在大数据环境下本质是错误的。大数据思维要求我们能够对数据有更多的思考,比如:是否可以服务(应用)于其他的应用场景?几乎可以武断地认为,任何来源的数据,除了满足当前用途外,一定有其他潜在的潜在应用价值,而这需要对数据本身、对各类应用具备全面、完备的认知度和敏感度,甚至是想象力,比如 Farecast 利用机票销售数据来预测未来的机票价格、谷歌重复使用搜索关键词来监测流感的传播、麦格雷戈博士用婴儿的生命体征来预测传染病的发生、莫里重新利用老船长的日志发现了洋流……

在数据价值评估的场景下谈论数据的价值，应该是其所有潜在用途的总和。对于数据的潜在价值评估，有三种最为常见的评估维度：数据的再利用价值、数据的可扩展价值和数据交叉组合的价值，以下逐一简单说明。

（1）数据的再利用价值

数据的再利用价值是指数据在完成原始的目标应用外，还能被用于其他的应用场景从而获得更多的收益。

数据的再利用价值需要原始数据的拥有者能够创新地发掘数据的潜在利用价值，而不是数据的初始价值目标达成后就将其束之高阁。一个典型例子是搜索关键词。消费者和搜索引擎之间的瞬时交互形成了一个网站和广告的广告列表，实现了那一刻的特定功能。乍看起来，这些数据在实现了搜索引擎的基本用途之后似乎变得一文不值。但是，以往的查询也可以变得非常有价值，比如谷歌利用用户搜索关键字能够预测 H1N1 流感、京东利用用户搜索关键词及在每个网页上停留的时间分析和预测消费者的偏好、英国央行通过搜索查询房地产的相关信息，更好地了解住房价格的升降情况、谷歌整理了一个版本的搜索词分析（公开供人们查询）并与西班牙第二大银行 BBVA 合作推出了实时经济指标以及旅游部门的业务预报服务……

数据的再利用价值对于那些收集或控制着大型数据集但目前却很少使用的机构来说是个好消息，比如那些线下运作的传统企业，他们或许正坐在尚未开发的信息喷泉上，有些企业可能已经收集了数据并使用过一次，且因为存储成本低而将其保存了下来，这些数据的再利用价值如果被发掘出，显然对各方都是大有裨益的。

由于在信息价值链中的特殊位置，有些公司可能会收集到大量的数据，但是他们并不急需使用也并不擅长再次利用这些数据。例如，移动运营商收集用户的话单信息、（每月）消费信息、手机的更换记录、用户的位置信息，所有这些数据都是移动运营商为了提供相应的移动通信服务必须（或者顺便）保存下来的，就这个意义而言，这些数据只具有狭窄的技术用途。而事实上，这些数据蕴含着每位消费者的朋友圈社交信息、行动偏好、消费偏好等，如果这些数据能够被再次利用，则变得更有价值。鉴于这些数据涉及个人隐私信息，出于法律监管的原因，或许还不能完全公开给第三方开发使用，但是，在运营商体系内的若干与数据再利用相关的创新项目已经展开。

（2）数据的可扩展价值

与数据再利用价值相对耦合的一个价值观是数据可扩展价值，所谓数据可扩展价值，指的是为了保障数据再利用价值的有效实现，在数据采集或者制造之初，就从顶层设计的角度规划好主动采集或者制造反映同一实体的数据并保障这一数据的可复用性。

促成数据再利用的方法之一是从一开始就设计好它的可扩展性，即相同数据集的多种用途的可能性，所以评估大数据的价值也需要考虑其可扩展价值。例如，出于安全防护的目的，

很多零售商在商店内安装监控摄像头。从安全防护的角度出发,这些设备的采购和安装无疑是一项纯粹的成本支出,但应当注意到的是,正是这样的视频监控采集到的数据能够实时反映在商店里购物的客户流和他们停留位置及停留时间,利用这些信息,零售商家可以设计店面的最佳布局并判断营销活动的有效性,同时生产厂家也可以利用这些信息判断自己的产品与竞品在消费者关注度上的差异,从而设计有针对性的营销策略。

这对于应用的提示或许是:在同一应用场景下,收集尽可能多的数据并在一开始就考虑到各种潜在的二次用途并使其具有扩展性,这对于增加数据的潜在价值无疑具有重要的意义。更多的利好消息或许是:在同一场景下,收集多个数据流或每个数据流中增加更多数据点的额外成本往往较低。因此,问题的关键是寻找“一分钱两分货”,即如果以某种方式收集的单一数据集有多种不同的用途,它就具有双重功能,这为数据再利用提供了有效支撑。

(3) 数据交叉组合价值

数据交叉组合价值是指(某渠道收集的)数据与(其他渠道收集的)数据组合在一起实现某个目标应用而体现出的数据价值。这里的价值或许是通过服务原始应用场景的目标应用获得的收益,又或许是在完成原始的目标应用外,再应用于其他的应用场景从而获得更多的收益。

有时,处于休眠状态的数据的价值(只能)通过与另一个截然不同的数据集结合(才能)释放出来,用新的方式混合这些数据,我们可以做出很有创意的东西来,比如“飓风-蛋挞”这个流行于大数据时代的故事,其实质就是将气象数据与超市的售货记录混合在一起发现的销售规律,类似的例子其实有很多,以每个人每天的活动数据来看:手机和每个基站的握手信息记录了携带手机的人每天的行动轨迹、铁路或者民航的乘车(机)信息反映了人的差旅信息、旅馆的登记记录反映了人的差旅入住信息……这些信息都散布于不同利益单位的服务器中,这些信息原本的价值是各个利益单位出于服务及管理而生成并在这一目标达成后作为历史数据存放在独立的数据库中,而如果将这些数据有效集成,则可以准确表征每个人的完整出行(记录),这不仅可以用于分析每个人的出行规律和习惯偏好,也是社会监管部门进行情报研判的重要数据。前者可以为商家精准营销提供决策依据,也可以为社会市场宏观分析提供数据支撑;后者则可以为案情研判、嫌疑人跟踪提供重要的信息情报。

事实上,“数据交叉复用”是大数据思维的重要特征,随着大数据理念不断“深入人心”,如何充分挖掘和发挥数据交叉组合价值成为大数据创新应用的重要策略。不过,不同数据源的数据都是以不同利益单位的自主性进行生成和收集的,因此如何充分发挥不同数据源数据的组合价值,一个重要的技术保障是进行有效关联,即在A数据源记录的某条数据和B数据源的某条数据表征的是同一个实体O。11.4.4节中给出了一个详细的示例说明,此处不再赘述。

随着数据存储成本的大幅下降,企业拥有了更强的经济能力来保存数据,并再次用于相同或类似的用途(或者向第三方开放),而大数据理念逐步“深入人心”,有条件的企业也愿

意在数据采集层次多做些投入,从而实现上述的数据再利用价值、可扩展价值及交叉组合价值。但是应当注意到的是,在大数据场景下讨论数据的价值,一个基本保障是数据应该具有“活性”。

众所周知,诸如亚马逊、淘宝、京东等电商平台会根据用户的购买记录、搜索记录(甚至包括在某个网页上停留的时间等)建立“用户-产品”的自动推荐模型,借此实现个性化的推荐服务。可以理解的是,随着时间的推移,大多数数据都会失去一部分基本用途。在这种情况下,继续依赖旧的数据不仅不能增加价值,实际上还会破坏新数据的价值。可以设想的例子是:如果这些电商平台利用若干年前的数据进行建模,这个模型的精准度是很难有实质的应用价值的。

因此用于建模的数据应该具有“活度”,这意味着两件事情:

1) 用于建模的数据要不断更新:去除(或保留)陈旧的历史数据的同时,增加每日产生的新数据。

2) 一旦训练数据更新,建模工作就要启动,这也意味着建模是持续进行的。

上述的第二点比较容易理解,从操作层面上看,第一点比较麻烦:如果一味保留所有的历史数据,将所有的历史数据和新近增加的数据汇聚在一起进行建模,已经对建模没有太大贡献的过于陈旧的数据势必会成为噪声。这意味着需要将陈旧的数据去除,而哪个时间段(以前的)数据是可以去除的陈旧数据呢?这是一个复杂难解问题。亚马逊等公司建立了复杂的模型来帮助自己分离有用和无用的数据。例如,如果客户浏览或购买了一本基于以往购买记录而推荐的书,电子商务公司就认为这项旧的购买记录仍然代表着客户的喜好。这样,他们就能够评价旧数据的有用性,并使模型的“折旧率”更具体。

不过应当注意的是,在很多实际应用场合中,陈旧的历史数据并不意味着可以像推荐模型建模那样直接去除,因为并非所有的数据都会贬值,所以许多公司提倡尽可能长时间地保存数据,即使监管部门或公众要求它们短时间内删除或隐匿这些信息。因为数据用于基本用途的价值会减少,但潜在价值却依然强大,以谷歌为例,谷歌一直拒绝将互联网协议地址从旧的搜索查询中完全删除,它只是在18个月后将最后四位数以隐匿搜索查询,因为这些数据对于改善未来的搜索体验依然有效。

基于上述的介绍和分析,我们会发现大数据本身的实际价值估计是一件极其困难的事情,诚然,数据的价值估算不再是将其基本用途简单地累加,但是如果数据的大部分价值都是潜在的,需要从未知的二次利用提取,那么人们目前尚不清楚应该如何估算它。这个难度或许类似于布莱克-舒尔斯期权定价理论出现前金融衍生品的定价或者专利等无形知识产权的估值,这或许意味着,如果不出意外,给数据的潜在价值贴上价格标签会给大数据行业带来无限商机。

类似于出版商从书籍、音乐或电影的获利中抽取一定比例,作为支付给作者和表演者的特许权使用费,在大数据时代,数据持有人倾向于从被提取的数据价值中抽取一定比例作为

报酬支付,而不是商定一个固定的数额。这样一来,各方都会努力使数据再利用的价值达到最大。然而,由于被许可人可能无法提取数据全部的潜在价值,因此数据持有人可能还会同时向第三方授权使用其数据,因而,在大数据时代,“数据滥交”可能会成为一种常态。

作为一种商业模式,这种通过许可使用权然后从中抽头的模式,在大数据应用场景下比较可行,不过应当注意的是,往往数据的持有者并不是主营数据业务(而且P2P的数据交易其效率也较低)。基于这样的原因,一些试图给数据定价的市场如雨后春笋般出现,2008年在冰岛成立的DataMarket向人们提供其他机构(如联合国、世界银行和欧盟统计局)的免费数据集,靠倒卖商业供应商(如市场研究公司)的数据来获利;总部设在得克萨斯州奥斯汀市的InfoChimps希望成为一个信息中间人,供第三方以免费或付费的方式共享他们的数据。谷歌的前员工吉尔·埃尔巴兹创办的Factual收集数据,然后制成数据库供需要者使用。微软也带着它的Windows Azure DataMarket登上了大数据的舞台,它的目标是专注高质量的数据,其方式和苹果公司监督其应用程序商店中的产品类似。微软假设,一位销售主管在准备Excel表格时可能还需要做一份公司内部数据和来自经济顾问的GDP增长预测的交叉表,那么她只要点击想要购买的数据,后者将瞬间出现在她的电脑屏幕上……

到目前为止,没有人知道估值模型将发挥怎样的作用。但可以肯定的是,经济正渐渐开始围绕数据形成,很多新玩家可以从中受益,而一些资深玩家则可能会找到令人惊讶的新生机。

11.4.2 数据质量评估

数据质量直接关乎数据的可用性,数据质量评估问题是信息化社会中固有的问题。而在大数据应场景下,多模式、多模态的数据呈爆发性势头增长,贯穿于数据生命周期的整个过程。我们如果难以保证数据来源以及数据分析过程中的数据质量,也就难以保证从数据中发现的知识和洞见的准确性,因此,数据质量监管是大数据应用中的重要一环。

大数据的价值最终体现在从数据中发现可以辅助生产实践的知识和洞见,这通过数据建模而来,显然,数据的真实性、一致性、精确性、完整性、时效性和实体同一性是保证数据建模有效的重要基础,上述的这些指标统称为数据的质量问题。在大数据时代,数据质量得到各界人士的普遍重视,一个朴素的原因就是在于:利用高质量数据,可以正确地做事以及做正确的事。

大数据时代对高质量数据的追求本质上是希望数据具有可用性。如前所述,数据的真实性、一致性、精确性、完整性、时效性和实体同一性是评估数据有用性的重要指标,分别描述如下:

1) 真实性指的是数据如实地反映对象实体,保证数据没有被恶意地、有倾向性地加工过,进而使得数据已经背离目标实体原本的样子。真实性是当前企业亟待考虑的重要维度,将促使他们利用数据融合和先进的数学方法进一步提升数据的质量,从而创造更高价值。出

于对真实性这个维度的重视, IBM 为大数据的 4V 特征添加了一个出于自身企业文化的第 5V 特征, 即 Veracity。

2) 一致性指的是数据集合中每个信息都不包含语义错误或相互矛盾的数据。

3) 精确性指的是数据集合中每个数据都能准确表述现实世界中的实体。特别注意一致的信息也可能含有误差而未必精确, 在许多应用领域, 信息精确性至关重要。

4) 完整性指的是数据集合中包含足够的数据来回答各种查询和支持各种计算。

5) 时效性指的是信息集合中每个信息(源)在数据的获取方面都是实时的、具有相当的“活度”。

6) 实体同一性指的是同一实体在各种数据源中的描述统一。

确保数据可用性是一项十分困难的任务, 而大数据的固有特点也使得确保大数据可用性将变得难上加难, 至少涉及以下几个理论技术难题(不限于):

(1) 高质量大数据获取与整合的理论和技術

在数据获取阶段把住质量关, 探索从物理信息系统等多数据源中有效获取高质量大数据的理论和方法, 研究高效数据过滤方法, 建立多模态大数据融合计算的理论和算法, 实现高质量数据获取和精准整合, 继而发现数据演变规律, 是确保信息可用性的重要前提。

(2) 完整的大数据可用性理论体系

建立大数据可用性的理论模型、大数据可用性的形式化系统和推理机制、大数据可用性评估理论和算法、大数据质量融合管理的理论和算法、大数据演化机理、大数据可用性所涉及的计算问题的复杂性理论和算法设计与分析的新方法, 借此建立完整的数据可用性理论体系, 为数据可用性研究奠定理论基础和量化标准。

(3) 数据错误自动检测与修复的理论和技術

提出大数据错误自动检测和修复问题的可计算性理论、大数据错误自动检测和修复问题的计算复杂性理论、大数据错误自动检测和修复方法的可行性理论、高效实用的大数据错误自动检测与修复算法, 借此在数据可用性理论体系基础上实现数据错误的自动检测和修复。

(4) 弱可用数据上近似计算的理论和技術

提出弱可用大数据近似计算的可行性理论、弱可用大数据近似计算问题的计算复杂性理论、弱可用大数据上近似计算结果的质量评估理论、弱可用大数据上的近似计算方法, 借此实现当数据中的错误不能彻底修复时(这些数据称为弱可用数据), 直接在弱可用数据上进行满足给定精度需求的近似计算。

(5) 弱可用数据上的知识发掘与演化的机理

提出源于弱可用数据的知识可用性评估理论与方法、数据可用性与知识可用性的相关性理论、弱可用大数据上知识发现的计算复杂性理论和算法设计与分析新方法、源于弱可用数据的知识校验与纠偏的理论和方法、源于弱可用数据的知识演变机理, 借此实现弱可用大数据上的知识发掘与演化机理分析。

综上所述, 针对大数据质量的评估在基础理论、算法和工程技术各层面都存在严峻的挑

战性研究问题。目前大数据可用性研究工作才刚刚开始,仅触及少数几个侧面,大量科学技术问题有待解决,向我们提出了新的挑战,也为我们提供了新的机遇。

11.4.3 平台价值评估

任何一个落地应用的大数据项目最终都是要面向最终用户的,其最终呈现和交互方式就是大数据平台(或者软件系统),因此大数据平台是否客观上满足用户的可用性需求(包括显式的或者隐式的)和主观上的用户体验需求以及满足的程度就成为用户评估一个大数据平台的重要依据。

用户一般是指直接与软件平台交互、以期完成某个任务的人,根据用户使用平台的频度,可以将用户分为主要用户(经常使用系统的人)、二级用户(偶尔使用系统的人)、三级用户(引入系统会影响到的以及影响购买的人)。这意味着软件平台的设计、实现与部署实施必须充分考虑这几类人的显式的或者隐式的需求。也有用当事人这个概念来描述进行一个系统设计与实现时需要考虑的需求来源,这里的当事人指的是被系统影响的而且对需求有直接或间接影响的个人或机构,比如甲方自身、乙方(开发团队)以及接受产品输出的人。

用户的可用性需求指的是目标系统需要完成的功能以及完成的功能在可行性、有效性、安全性、通用性、易学性、易记性等方面的客观诉求;用户体验需求指的是目标系统能够多大程度上引起用户积极反应,让用户感觉轻松、舒适并能从中获得享受。前者需要通过需求工程对用户场景及目标进行定量或定性的收集、分析和归纳;后者则需要在对人性因素的理解基础上进行有针对性的设计与实现。

评估一个大数据平台的维度一般包括(不限于):

1) 数据获取能力:包括数据收集能力、数据融合能力、数据预处理能力等,一般而言,这方面的能力体现往往涉及一些非技术因素,比如商务合作能力以及政策理解、法律遵循,而且在进行具体的数据采集、融合和预处理时,往往带有一定的目标指向。

2) 平台构建能力:包括数据存储能力、数据计算能力、数据分享能力,这方面的能力直接影响甚至决定了以此为后续目标应用系统的构建和部署,因此这方面能力的实现往往需要基于合适的计算模型、架构及服务模式,当然也需要必要的硬件存储与计算设备作为最基础的保障。

3) 数据应用能力:包括数据建模能力、数据呈现能力、数据管理能力,对于围绕具体垂直应用目标开展的大数据项目,数据应用能力往往与业务目标紧密相关;如果大数据项目是面向普适应用的,则还需要考虑其能多大程度上支持二次开发。

以下详细介绍每个评估维度的具体要求:

(1) 数据收集能力

“数据为王”是大数据平台建设者的共识,如何保障大数据平台的数据“广”度和“全”度是大数据平台得以有效部署实施的重要基础。鉴于数据来源广泛,因此能否从众多数据源

中获取数据是评估大数据平台数据收集能力高低的重要指标。ETL 工具和互联网网络爬虫工具是两类典型的数据收集工具,前者用于从自营数据平台或者(通过某种合作协议)其他利益数据平台获得既有的数据,这类数据的格式和规范往往是以先验经验的形式隐式存在;后者往往专注于从互联网上收集相关数据。无论数据的来源如何,尤其是针对互联网数据源,需要并行考虑的问题是数据源的动态更新,这意味着大数据平台除了能够从配置的数据源中获取数据外,还要能够对动态更新的数据源有自适应能力,这些都是考量数据收集能力的重要维度。

(2) 数据融合能力

多样性是大数据的一个重要特点,大数据的多样性特点包括数据源的多样性以及数据类型的多样性,而大数据应用的特点在于通过建模从数据中提取出模式(数据动态变化时,模式也在动态变化),这就意味着要想处理大数据,首先必须对所需数据源的数据进行抽取和集成,鉴于数据的异构性是大数据的普适特点,因此能否有效应付大数据场景下异构性的典型特点,是评估大数据平台数据融合能力高低的重要指标。

传统的数据集成中也会面对数据异构的问题,但是在大数据场景下异构性的典型特点在于:①数据类型从以结构化数据为主转向结构化、半结构化、非结构化三者的混杂。②数据产生方式的多样性带来的数据源变化,传统的电子数据主要产生于服务器或者是个人电脑,这些设备位置相对固定,而随着移动终端的快速发展,手机、平板电脑、GPS 等产生的数据量呈现爆炸式增长,且产生的数据带有很明显的时空特性。③数据存储方式的变化,传统数据主要存储在关系数据库中,但越来越多的数据开始采用新的数据存储方式(比如 NoSQL)来应对数据爆炸,这就必然要求在集成的过程中进行数据转换,而这种转换的过程是非常复杂和难以管理的。

(3) 数据预处理能力

数据量大是大数据的另一个重要特征,而数据量级的增加往往意味着数据质量的下降,因此在数据分析之前必须进行数据清洗等预处理工作,但是预处理如此量级的数据对于机器硬件以及算法性能都是严峻的考验,因此数据预处理能力是评估大数据平台的一个重要维度。

如果在集成的过程中仅仅简单地将所有数据聚集在一起而不作任何数据清洗,会使得过多的无用数据干扰后续的数据分析。另一方面,大数据时代的数据清洗必须尤为谨慎,因为相对细微的有用信息混杂在庞大的数据集中,如果信息清洗的粒度过细,很容易将有用的信息过滤掉。清洗粒度过粗又无法达到真正的清洗效果,因此在质与量之间需要进行仔细的考量和权衡。

(4) 数据存储能力

数据一定是以集中或者分布的形式存储在大数据平台中的,因此一个大数据平台的数据存储能力是评估大数据平台的重要指标。大数据应用场景下,数据的“大”不仅体现在数据存储量上的“大”,还要考虑后续应用对数据存取吞吐率的“大”。这意味着,在评估大数据

平台的数据存储能力时,不仅要在静态上考察存储容量,还要考察面向后续应用的并发“存”和“取”的效率支撑能力。

(5) 数据计算能力

计算是大数据平台向用户提供各类服务的载体,为了向用户提供各种(功能)服务,需要的前置工作包括数据采集、预处理、建模分析、检索等,每一个环节都需要有高性能计算保障,否则无法响应大数据的挑战,也无法提供满意的用户体验。计算能力的提升涉及硬件和软件两个维度,在硬件能力给定的情况下,软件架构以及软件的实现是保障计算能力的重要基础,这也是评估大数据平台数据计算能力的重要指标。

(6) 数据分享能力

大数据的一个重要思维是开放和合作,这不仅意味着大数据项目的建设者可以(有限)开放数据的访问以获得相应的社会价值回报,更重要的是,数据本身即可成为一种资产,为第三方(软件开发商)提供数据支撑从而带来收益。而这都需要在数据层提供高效的访问接口,使得数据的潜在需求者能够便捷地获取数据服务。

(7) 数据建模能力

数据分析是整个大数据平台的核心,从异构数据源抽取和集成的数据构成了数据分析的原始数据集,围绕不同的目标应用需求,从这些数据中选择全部或部分进行建模,建模的好坏直接影响从数据中获取的知识和洞见以及满足目标应用的匹配度,因此数据建模能力评估是大数据平台的重要指标,可参照的评估维度有(不限于):

1) 数据建模的实时性。大数据的“活度”要求数据建模必须实时跟随(响应)数据的“动态(变化)”,否则从陈旧数据中建立的模型价值会逐步衰减(极端情况下会产生错误的结果),因此数据建模的实时性是大数据应用的基本需求。

2) 动态变化环境中索引的设计能力。关系数据库中的索引能够加快查询速度,但是传统的数据管理中模式基本不会发生变化,因此在其上构建索引主要考虑的是索引创建、更新等方面的效率。大数据时代的数据模式随着数据量的不断变化可能会处于不断的变化之中,这就要求索引结构设计简单、高效,能够在数据模式发生变化时很快地进行调整适应,这是大数据时代的一个挑战,也是评估大数据平台先进性的一个重要指标。

3) 对先验知识的响应能力。对于任何一个目标应用场景的数据建模,往往离不开领域知识的支撑(领域知识一般以显式的规则、策略、思维或者隐式的先验经验体现)。仅在数据层面进行的普适数据建模往往会使得建模的结果忽略领域的个性化特征,从而使得建模难度加大或者建模效果欠佳,因此,大数据分析师能够多大程度上理解和包容领域知识直接决定了大数据平台(面向垂直目标应用)的应用效果。

(8) 数据呈现能力

不仅仅是因为提供友好人机交互需要可视化,大数据应用场景下,数据可视化的需求更为迫切,可能的原因是:①数据的复杂性需要大数据平台能够以图形化的方式帮助人们更好地理解数据。②数据分析流程的复杂性需要大数据平台能够以图形化的方式展示和配置分析

流程以帮助用户更好地理解、监管和回溯分析流程。③数据分析结果的多维度也需要大数据平台提供图形化的方式以便于用户从多个维度使用和评估分析结果。因此数据呈现能力是评估大数据平台的一个重要维度,当然图形化仅是数据表示的一种方式,或许还有其他的表示方式,此处不一一赘述。

(9) 数据管理能力

大数据平台赖以存在的基础是数据,在大数据平台下,所有的基础素材都是从不同数据源采集到的数据,因此数据管理贯穿于整个大数据分析流程的重要环节,必须有稳定、高效、合理的数据管理方法、策略以支撑大数据平台的稳定运行,其涉及的内容也贯穿于“数据→价值”的整个流程中。

(10) 其他一些考量因素

作为一个面向用户应用的软件系统,大数据平台除了需要在上述的价值维度上有所考虑外,仅仅作为一个软件平台,可以考量的因素还有很多,以下简单罗列:

1) 平台易用性。易用性是贯穿整个大数据流程的重要指标,且随着大数据应用的不断深入,这个指标的重要性愈发突出。应用需求使得很多行业都开始进行大数据项目的建设,但是这些行业的绝大部分从业者都不是数据分析的专家,在复杂的大数据工具面前,他们只是初级的使用者,复杂的分析过程和难以理解的分析结果限制了他们从大数据中获取知识。这意味着,在大数据应用场景下,设计、实现易学、易用的软件工具至关重要。

2) 平台的能耗。大数据时代的来临使得传统构建计算中心的思维转为构建数据中心的思维,随着数据规模以及人们对数据需求和期望的不断提高,数据中心的规模也处于不断增长过程中,而数据中心所产生的能耗问题在全社会倡导绿色与节能的大趋势下逐步成为大数据项目建设中的一个社会问题(其实也是技术问题),因此,在大数据平台的评估中,能耗也会成为一个重要的评价指标。

在大数据管理系统中,能耗主要由两大部分组成:硬件能耗和软件能耗,二者之中又以硬件能耗为主。《纽约时报》和麦肯锡经过一年的联合调查,发现 Google 数据中心年耗电量约为 300 万瓦,而 Facebook 则在 60 万瓦左右。最令人惊讶的是在这些巨大的能耗中,只有 6%~12% 的能量被用来响应用户的查询并进行计算。绝大部分的电能用以确保服务器处于闲置状态,以应对突如其来的网络流量高峰,这种类型的功耗最高可以占到数据中心所有能耗的 80%。从目前的研究成果来看,采用新型低功耗硬件以及使用可再生的新能源(如太阳能、风能)是两种主流的研究思路。

11.4.4 应用提示

一个关于大数据的描述是“大数据是数据本身以及数据采集的工具、平台和分析系统的总称”,就这个意义而言,针对大数据的评估至少应该从数据本身以及以数据为中心、应用为目标的服务平台这两个维度展开。

1. 数据层面

在“数据为王”的大数据时代，“数据有价值”及“数据即资产”是大数据时代的共识。大数据产业链中的各个业务形态（专注于数据收集与整合的数据型公司、基于数据分析及计算架构的技术型公司、基于数据驱动型的面向具体应用开发的思维型公司）都直接或间接地与数据相关。

（1）数据质量

如前所述，“数据的真实性”是评估数据质量好坏的一个重要维度，需要注意的是：“数据的真实性”这个评估维度是与目标应用相关的动态属性，这意味着只有有了明确的应用目标（及应用场景），才能确定这个目标应用（及应用场景）相关的实体对象是哪些，也只有确定了实体对象，才能定性地表征这个数据反映的是真的还是假的。

仍然以 11.2.2 节中提及的软文营销为例，软文营销一直是互联网营销中的一个重要内容，其实质就是由商家本身或者雇佣写手编撰、发布和转发对商家产品有利的新闻、故事、心得、评述等，而此类软文往往以隐性投放为主（与一般的显性广告相反），消费者往往并不知道此类“信息”的真与假，往往在潜移默化中逐步地、不设防地接受商家的“灌输”。显然，以消费者为中心，为消费者梳理各类产品信息或者金融情报时，此类（往往充斥互联网）软文（广告）不是反映该企业产品形象的真实数据，应该在预处理中过滤；而如果从竞争对手分析的角度来看，这些软文（广告）恰恰是反映竞争对手营销策略、能力和市场边界的真实数据。

时效性关注的是数据的“活度”，这意味着在进行数据的收集和整合时，要有意识地选择数据持续更新的数据源，这或许是不得已而为之的选择，根本的原因在于：隐藏在数据背后的模式（事先并不可知）必须通过数据建模来进行“构建”，而通过历史数据构建的“模式”（模型 A）未必是操纵数据的“上帝意图”（模型 B），在这种假设下，唯有利用（模型 B 操纵生成的）更新的数据进行建模，才有可能使数据建模构建的模型 A 更贴近那种“上帝意图”随着时间的推演而反映在数据中的“偏好”。

楚人有涉江者，其剑自舟中坠于水，遽契其舟，曰：“是吾剑之所从坠。”舟止，从其所契者入水求之。舟已行矣，而剑不行，求剑若此，不亦惑乎！

《吕氏春秋·察今》中记载的这则“刻舟求剑”的故事是告诉人们：世界上的事物，总是在不断地发展变化，人们想问题、办事情，都应当考虑到这种变化，才能适应这种变化的需要。这对大数据分析师的提示在于，隐藏在数据背后的模式随着时间的变化往往会有所变化，因此必须通过更多的反映这种模式变化的数据来进行持续的建模，才有可能及时跟进模式的演化。

实体同一性关注的是数据的“靶向性”，即不同数据源在反映同一实体上的表示统一，实际操作中的最大难度或许是不同数据源的实体映射关系的发现，这往往需要专门的技术或者策略（甚至包括一些政策或法律的支持）协同进行。

从互联网中采集和整合目标客户群的偏好是商家进行互联网营销的重要策略,而如果能够获得个体的立体画像(比如行为习惯、消费偏好)显然能够有助于提高“推荐→支付”的转化率。以普通人群常用的电商或社交平台为例,在淘宝、京东、苏宁易购上购物(并发表评论),在QQ、微信或者微博上发表见闻或者感想,如果能够以“用户”为中心收集各个平台的数据,显然对于个性化营销大有裨益。但是用户在各个平台上的用户名往往是不一样的,即一个(物理)用户在不同的平台上拥有不同的(虚拟)用户名 nameA、nameB、nameC……如何将这些虚拟用户名加以整理并与一个具体的(物理)用户对应在实际操作中难度极大,主要体现在:

1) 用户在淘宝、京东、苏宁易购等电商平台上的销售型数据是掌握在各个平台商手中,而QQ、微信的社交型数据是掌握在腾讯手中,旁人无法获得完整的数据(这或许可以解释电商型公司和社交平台型公司战略合作的缘故)。

2) 即便能够将这些数据收集在一起,也需要有技术手段识别出此平台中的 nameA 与彼平台上的 nameB 是同一个人。在进行用户名注册的时候会绑定手机号、用户在使用手机登录系统时及时记录下对应的 MAC 地址等操作,这或许为虚拟身份的统一奠定了数据基础,但即便如此,仍然需要考虑的是手机号或者 MAC 地址是否与一个物理的人一一对应?

我国出台相应法律法规要求手机号必须与身份证对应,或许在法律层面上能够将手机号(这个虚拟身份)和物理身份统一起来,但显然这不是为电商公司及社交平台型公司服务的,而且这个数据是掌握在电信运营商手中,又是另外一个数据源,而这些数据源的融合显然不仅仅是技术问题。

一致性、精确性、完整性关注的是数据的语义“厚度”,即数据在反映物理实体语义上的一致性、完备性和可信性方面的程度有多大,前面已经多有表述,此处不再赘述,对于应用的提示在于:在数据的收集阶段,尽可能收集具有结构语义的富信息,同时确保信息的一致性和完备性。

(2) 数据价值

数据一定是有价值的,这是人们对大数据极其热衷的最重要原因,从技术流的角度而言,如何从数据中发现有助于未来实践的知识自然是大数据的重要价值体现,而在该过程中产生的若干衍生品其价值同样重要,具体包括:

1) 数据本身就有价值,这意味着,在大数据这个产业链中,能够将不同数据源的数据收集(包括采购)到一起,就可以作为资产卖给第三方,当然,如果能够对数据进行一些加工(比如去重、规约、结构化),则可以以更高的价格进行销售。与此直接耦合的另外一个问题就是数据的定价问题,鉴于数据的原始价值、交叉复用价值以及可扩展的价值并存的事实,将数据以服务的形式提供,以服务的质量和体量进行定价或许是有效的可选策略之一。

对于以数据业务为主营目标的公司,其业务模式可以简单地梳理为:产品经理罗列和遴选数据源(数据在哪里?)、技术团队设法获得相关数据源的数据并加以预处理(如何取?)、

营销经理寻求数据需求方（用在哪里？）。在实际操作中，这三个环节是彼此耦合的，而且各个环节均会涉及技术及非技术因素的影响：产品经理在梳理数据源的时候往往需要考虑潜在的应用场景、数据获取的技术成本及商务成本；技术团队在获取数据的时候往往需要考虑数据的组织、存取和管理（如何组织、如何存、存在哪里、是否以及如何加工、如何取）；营销经理承担着整个数据业务的利润获得，显然岗位职责尤为重大，其面临的第三方需求往往有两类，一是直接找出数据的潜在需求方（甲方），直接将既有数据以某种方式提供给甲方（一般还会根据甲方的需求扩展采集更多的数据源数据），另外一个来源于甲方的需求场景，根据既定的甲方的目标应用及既有数据，提供完整的数据解决方案，这种需求往往在实际数据运营过程中占有很大的比重。这意味着作为一个数据型公司，往往因为甲方的此类需求使得自身作为纯粹数据型公司的定位变得模糊，除非委托另外的第三方进行方案的设计与部署，而这又牵涉到与第三方的商务合作。

2) 从数据中提炼出一些反映目标实体的标签化的信息对于很多的应用场景而言，就直接有用。比如，为商家提供具有某些标签的客户群、区域。数据的拥有者利用数据对实体进行标签化事实上是对数据的“精加工”，能够大范围缩减第三方的工作量的同时，获得比销售原始数据更高的利润，而且对自有数据资产的版权有更高层次的保护。

在很多应用场景下，数据的实际需求方实际需要的并不是数据本身，而是在数据基础上凝练的情报（或者信息）。比如对于广告商（商家）而言，需要的是潜在的广告受众群在哪里？对于面向即刻消费的便利店而言，需要的是各个地段的人流量以及这些人流中的目标客户群的分布有哪些？对于案源分析的情报研判人员而言，需要的是与案源有关的线索……所有这些潜在需求都表明数据的提供方应该在原始数据及预处理加工的数据基础之上进行一些语义信息的提取，通过技术手段对原始数据或预处理加工的数据语义标签化，然后将这些标签提供给潜在的需求方。这些标签针对的对象包括数据（数据可信度及权威度、数据内容摘要、数据使用率、数据重复度等）、物理实体（数据相关的人、物、地点、事件等，“事件”往往是需要通过数据发现的）等，这些标签所覆盖的维度包括：通用标签（描述物理实体、事件或数据本身的关键热词、静态属性等）、情感标签（描述物理实体、事件或数据本身的情感倾向）、高级语义标签（描述物理实体、事件或数据本身的风险度、健康度、价值度等）及领域专用标签（与目标应用领域直接相关的，比如“手机发烧友”“购物狂”“夜猫子”“户外爱好者”）。

3) 利用自有技术对数据进行分析加工得到的数据模型，可以以计算服务的形式提供给第三方，也是大数据产业链中的重要一环（对应于技术型公司）。在大数据实践中，数据模型往往是持续不断地改进的，这意味着大数据项目的建设需要持续的运维，而这种运维相比传统的软件工程中的系统运维要复杂得多，传统的系统运维往往关注软件系统的容错性、适应性、完善性，而大数据运维除了要满足上述的基础需求外，仅就数据建模而言，还需要持续迭代。

从数据中发现知识和洞见一直被认为是大数据的重要价值，而这种知识和洞见往往是通

过数据建模来体现,鉴于大数据平台的应用方往往并不是数据建模的擅长者,即便是大数据项目的实施方,也未必对所有的建模方法有十足的攻关能力。因此,在大数据时代涌现了大量的以技术研发为核心竞争力的公司,这些公司专门针对大数据应用中的若干技术细节(通用的或者专门服务于某一应用领域的)提供解决方案。这不是关键,关键在于:即便是大数据项目部署实施得以运营,这些建模的工作也要持续进行(包括迭代改进和完善),这意味着,在大数据场景下,大数据部署实施以后的运维往往更为重要。因此,在大数据时代,“大数据平台+人力外包”的商业模式甚为流行,其基本的做法是:大数据平台实施方根据甲方的需求构建完大数据平台投入运行后,平台实施方继续投入人力进行运行过程中的迭代开发和维护,特别注意的是,这种运维有别于大数据平台实施后甲方投入人力进行的业务层面及商务层面上的运维。

4) 能够从数据中研判、挖掘出这个领域的知识,从而为领域专家理解和分析这个领域的公理系统、定理系统提供辅助决策。这对于应用的提示在于:进行大数据项目建设的时候,一个隐式需求是为领域用户(一般是专家)提供数据后研究的功能,以便为领域专家进行领域研究提供数据基础和平台支撑。

从生产实践中获得公认的并用于指导未来生产实践的知识是人类文明进程中亘古不变的活动,或许也是人之所以成为地球主宰并得以持续发展的重要原因。在大数据场景下,人们生产实践的现象、状态或者结果均以数据的形式存放在数据库中,如何从这些数据中获得指导未来实践的知识应该是大数据时代的重要任务,这也是大数据(技术)在各个领域得到广泛认可和推广的主要原因。不过应当注意,通常意义上的大数据项目实施中的数据建模关注的是某个时间段内的数据模型的即时效用,即当前的模型仅服务于短期的目标(随着数据的持续更新,模型也要持续改进),这也是上述大数据持续运维的本质原因。就这个意义而言,此类数据模型还没有反映和表示可以称得上领域知识的能力,显然,这类领域知识是极其重要的,需要领域专家从领域学科和方法论的角度加以研究和推进,作为大数据平台能够为领域专家的这类需求提供尽可能完备的数据服务和计算服务。因此,大凡一个大数据项目的建设,都应该把这个需求作为一个官方的隐式需求纳入到大数据项目建设中。

2. 平台层面

(1) 平台本身价值

对于任何一个大数据平台,其逻辑上包括数据平台(专注于数据的收集、存取与管理)、计算平台(专注于数据处理、分析与建模)、应用平台(专注于目标应用的子功能响应)及综合服务平台(专注于人机交互及综合管理),从评估的角度出发,可以从数据获取、平台构建及数据应用几个维度对整个大数据平台进行定性及定量评估。

大数据平台的首要价值体现在满足和匹配目标应用的程度上,“大数据、小应用”是大数据时代的一个典型特征,这意味着:①大数据平台是以数据驱动的应用开发,数据是维系大数据平台应用的根本。②针对必然复杂的业务逻辑应用的响应应该是一系列的小应用(系统)

的有序组合。这对于应用的提示在于：在围绕目标应用进行数据采集与整合的基础上，应该对业务应用进行梳理和重组，以实现“小应用”的最大复用，在实际操作中，往往会涉及职能部门的业务重构。

以基于C2B的柔性制造厂（商）为例，从职能的分工划分，制造厂（商）或许会设立专门的产品设计部、采购部、市场部等，在数据采集平台完备的基础上，这些部门的需求响应事实上可以归结为很多小应用的有序组合。

以产品设计部为例，其业务职能在于回答：做什么？为什么做？如何做？为了有效回答这些问题，需要的应用支撑包括（不限于）：①根据“产品分析系统”，目前用户需要此产品。②根据“市场分析系统”，产品值得做。③根据“竞品分析系统”，用户希望产品应当具备属性（竞品缺陷分析及负面舆情分析），设计部门应该匹配需求。④根据“竞品分析系统”，市场前景甚好，时不我待。⑤根据“舆情分析系统”，现有用户反馈我司产品不尽人意，需要改进。

而对于市场部门而言，其业务职能在于回答公司的网络口碑如何？有无大的问题？竞争对手有哪些？他们怎么样？市场大环境怎么样？有无不适的新闻？为了有效回答这些问题，需要的应用支撑包括（不限于）：①根据“舆情分析系统”，提示舆情摘要并能实时预警。②根据“市场分析系统”，提供行业新闻报告摘要，提示政策、市场风险。③根据“竞品分析系统”，提示竞争对手现状（新闻、营销活动），提示需要做哪些相应响应。

基于上述两个部门的简单分析，可以明显看到两个不同职能的部门在“舆情分析系统”“竞品分析系统”“市场分析系统”是有需求交集的，这意味着在业务梳理的时候应该使响应需求的应用粒度尽可能更小些，以便更复杂业务应用的有序组合。或许这也意味着在大数据时代讲究“数据交叉复用”的同时，也要讲究“小应用的交叉复用”。

（2）平台扩展价值

大多数情况下，大数据平台的构建都有其最原始的目标导向，比如面向某个具体的应用需求。在当前各个利益主体都在建设大数据平台的现状下，一个突出的隐含问题在于：如果各个大数据平台不能做到彼此共享和复用，那么每个大数据平台事实上又是一个一个“数据（信息）孤岛”，因此，如何确保每个大数据平台的可复用是大数据平台（扩展）价值的重要体现。因此，大凡一个大数据平台部署实施，甚至在大数据平台建设伊始，就需要考虑该大数据平台的潜在扩展可能，至少体现在两个层次：①当前平台的数据有多大的开放度允许第三方的复用，借此实现数据的交叉复用价值和可扩展价值。②当前平台的数据建模和分析方法有多大的开放度允许第三方复用，借此实现计算服务的增值。

以“互联网+购物”应用场景为例，生产厂（商）通过线上线下融合实现产品从“工厂→消费者”的扭转，此过程中涉及的环节包括物流、仓储、实体卖场、电商卖家等。从生产厂（商）的角度而言，出于对市场行情、产品舆情的把握，生产厂商会有意愿进行产品舆情大数据平台的开发（此处仅以产品舆情导向为例）；从政府监管部门的角度而言，以“市场监管、消费维权”的工商部门为例，也有意愿构建面向整个市场的产品舆情大数据平台；消费者出

于对消费产品的关注和反馈,也会具有获得产品舆情需求,而这块需求往往散布于一些电商网站、搜索引擎或者地方论坛上……事实上,仅以产品舆情这一个需求为例,就有至少这么三个角色部门会构建类似的大数据平台,而事实情况也正是这样进行的。显然,这三类平台在目标导向、数据源甚至分析手段等方面均具有很大的相似性。当然,不同的部门出于自己的角色地位的不同,在数据源、分析手段上会有差异,而这种差异性也意味着这三者事实上是有很大的互补性,因此,从集约发展的角度而言,这三个大数据平台事实上应该进行数据层、应用层的分享以实现最终的共同目标。

11.5 本章小结

“Value”是大数据的一个重要特征,或许也正是因为“有价值”,才引发了“政产学研商用”各界对大数据的普遍关注、跟踪甚至追捧。本章尝试梳理的几个要点在于:

1) 应用为本:大数据的价值在于面向具体应用的落地部署和实施,以及在实现这一目标价值的同时,通过多边协同与协作而不断涌现出的理论创新、技术创新、应用模式创新以及商业模式创新等。

2) 数据为王:数据是一种资产,可以进行买卖,而实际上可以买卖的不仅限于原始数据本身,还包括进行预处理后的数据、从数据中凝练的信息以及从信息中发现的模型和知识,越是经过加工提炼的数据,其价值度越大,其扩展应用的灵活性越大。

3) 数据交叉复用:数据的价值除了满足最原始的应用目标外,还具有可扩展价值,可扩展价值需要通过数据交叉复用得以体现和实现,当然更需要有足够的敏感度和智慧去发现潜在的可能。

4) 大数据小应用:物理上被分开的业务部门在业务需求的逻辑上往往具有很大的耦合性,因此在进行业务梳理和重组的同时,应该以更小的粒度开发数据驱动的小应用,借此为小应用的有序组合提供可灵活复用的基础。

5) 共享开放是主旋律:各个大数据平台的建设方都应该有更多的包容精神实现自身平台的共享和开放,这不仅会带来更多的利润源,也是破解各个大数据平台又将成为孤岛这一魔咒的法门。

6) 结构优化是大趋势:无论是闭环还是开环,其目标都是服务于整个生态系统中的某一个环节的应用,因此服务于整个生态系统是大数据平台的终极目标,需要多边的有序集成、协同和协作。

“天生我材必有用,千金散尽还复来”(摘自《将进酒》)表述了诗仙李白自信为人的自我价值,也流露出怀才不遇和渴望用世的积极思想感情。从正能量的角度来看,这或许也是在说明:造物主成就的每一个人(实体)都有其立于世间的价值,关键是有没有相应的伯乐能够发觉并实现此价值。

在大数据应用场景下,这或许也是大数据平台建设者应该有的一种情怀。

本章参考文献

- [1] Assunção M D, Calheiros R N, Bianchi S, et al. Big Data Computing and Clouds: Trends and Future Directions [J]. Journal of Parallel and Distributed Computing, 2015: 79: 3-15.
- [2] Chen H, Chung W, Xu J J, et al. Crime Data Mining: A General Framework and Some Examples [J]. Computer, 2004, 37(4): 50-56.
- [3] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting Influenza Epidemics Using Search Engine Query Data [J]. Nature, 2009, 457(7232): 1012-1014.
- [4] Jones N D. Computability Theory: An Introduction [M]. Salt Lake City: Academic Press, 2014.
- [5] Koomey J. Growth in Data Center Electricity Use 2005 to 2010 [R]. Analytical Press, 2011.
- [6] Lynch C. Big Data: How Do Your Data Grow? [J]. Nature, 2008, 455(7209): 28-29.
- [7] Mantin B, Koo B. Dynamic Price Dispersion in Airline Markets [J]. Transportation Research Part E: Logistics and Transportation Review, 2009, 45(6): 1020-1029.
- [8] Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think [M]. Boston: Houghton Mifflin Harcourt, 2013.
- [9] Nath S V. Crime Pattern Detection Using Data Mining [C]. Web Intelligence and Intelligent Agent Technology Workshops, 2006: 41-44.
- [10] Paliathanasis A, Krishnakumar K, Tamizhmani K M, et al. Lie Symmetry Analysis of the Black-Scholes-Merton Model for European Options with Stochastic Volatility [J]. arXiv preprint, 2015.
- [11] Ruiming T, Huayu W, Zhifeng B, et al. The Price is Right: Models and Algorithms for Pricing Data [C]. 24th International Conference on Database and Expert Systems Applications, 2013: 380-394.
- [12] Shin D H, Choi M J. Ecological Views of Big Data: Perspectives and Issues [J]. Telematics and Informatics, 2015, 32(2): 311-320.
- [13] 伯纳德·利奥托德, 马克·哈蒙德. 大数据与商业模式变革: 从信息到知识, 再到利润 [M]. 郑晓舟, 胡睿, 胡云超, 译. 北京: 电子工业出版社, 2015.
- [14] 刘显敏. XML 数据实体同一性相关技术的研究 [D]. 哈尔滨工业大学, 2013.

大数据思维

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的杨骏元、张弛、王陆霞、汤北亮、方贺贺等几位同学的协助，在此表示深深的谢意。

12.1 引言

上古伏羲氏时代，洛阳东北孟津县境内的黄河中浮出龙马，背负“河图”，献给伏羲，伏羲依此而演成先天八卦并创立六十四卦历法……（所谓“河出图，洛出书，圣人则之”，摘自《易·系辞上》）

公元前 1054 年，82 岁的周文王被囚于羑里监狱，作后天八卦并赋卦辞；公元前 1038 年，周武王去世，太子诵继位（成王），周公（周文王四子）辅佐，制礼作乐，修八卦，增加爻辞……

公元前 551 年，孔子出生，晚而喜《易》，序《彖》、《系》、《象》、《说卦》、《文言》……（摘自《史记》）

1679 年 3 月 15 日，莱布尼兹发表论文《二进制的算术》……

周易八卦和二进制是否有关系是很多人茶余饭后的谈资，原因或许在于：①二者太像了。②周易八卦的出现远早于二进制的诞生。往往因为自身的价值偏好或心理习惯，人们倾向于建立起事与事之间此因彼果的关联（比如有传言“莱布尼茨受《周易》的影响创造二进制并用于计算机”）。显然，如果没有更多的史料和证据的话，这种关联连休谟所谓的“经常联结”都谈不上（此方面内容参见 9.4.2 节）。而经过考证的事实是：“莱布尼茨先发明了二进制，后来才看到传教士带回的宋代学者重新编排的《周易》八卦，并发现八卦可以用他的二进制来解释”（摘自郭书春所著《古代世界数学泰斗刘徽》）。

被莱布尼茨誉为“二进制的中国版”的《周易》发明于几千年前，但是聪明的中国古代人并没有将八卦中无意涉及的理念有意识地发展为类似二进制这样的理论体系，而是将其向

历史、人文、哲学等方向发展，并在漫长的历史中逐渐演变为一本“智慧之书”。这或许是中国人的大智慧，也是中国式的思维方式使然。

事实上，即便是被公认为二进制的发明者——莱布尼茨也不是这种记数法的最早发现者。17世纪初，英国数学家哈利奥特在他未发表的手稿中提到过类似的记数法；1670年，卡瓦利埃里又重复了这一发现。莱布尼茨的伟大之处在于不仅发现了二进制，还对二进制进行了充分的讨论，并建立了二进制的表示及运算。正是因为莱布尼茨的大力提倡和阐述，二进制才引起世人关注，并因为其在计算机实现方面的天然优势而大行天下。显然，这是莱布尼茨的智慧，或许也是西方的思维方式使然。

恰如《周易》和二进制，针对同一现象，不同的人从不同的思维角度出发，可以引申和发展出不同的创意。甚至同一个人，从不同的思维角度也可以产生和引发不同的决策和行动，或许这也是“大千世界、芸芸众生相”的根本原因，这是思维的力量。

思维是人用头脑进行逻辑推导的属性、能力和过程，是人脑对客观现实概括的和间接的反映，它反映的是事物的本质和事物间内在的、必然的、规律性的联系，属于理性认识。具体可分为逻辑思维、形象思维、知觉思维、顿悟等，其基本过程无非都是分析与综合、比较与分类以及抽象与概括。而被冠之为“某某思维”的思维，其实指的是人们面对某个具体的现象或事物，应该具备的思维方式。比如百度创始人李彦宏首次提到“互联网思维”这个词时说，“我们这些企业家们今后要有互联网思维，可能你做的事情不是互联网，但你的思维方式要逐渐像互联网的方式去想问题”。伴随着各界对此概念的认同，“互联网思维”也演化出不同的版本。比如小米的雷军就将互联网思维进一步提炼为“专注、极致、口碑、快”，这似乎已经成为互联网营销的圣经；而腾讯眼中的互联网思维则是“用户、简约、极致、迭代、流量”，依此扩展出社会化思维、大数据思维、平台思维等，并提出了“互联网+”这个新概念。

那么在大数据环境下，我们应该如何用大数据的思维方式思考问题呢？大数据至少涉及数据、技术、应用、商务等层面的方方面面：

- 1) 数据层面，面对及必须要考虑的问题至少有：是否有足够大足够多的数据？数据在哪里？数据类型和分布如何？数据的活度如何？
- 2) 技术层面，面对及必须要考虑的问题至少有：如何获得数据？如何存取数据？如何分析数据？如何搭建系统？系统架构如何？如何部署实施？
- 3) 应用层面，面对及必须要考虑的问题至少有：目标应用在哪里？应用模式怎样？基础环境如何？如何提供服务？业务梳理清晰与否？渠道有哪些？目标受众是谁？
- 4) 商务层面，面对及必须要考虑的问题至少有：商业模式如何？法律法务支撑环境如何？政策环境如何？如何营运运维？是否有足够的基础保障？

上述的这些问题或许远不够全面，但核心的问题是在不同的层面我们应该用怎样的思维方式去匹配大数据的特点。

本章尝试从数据层、分析层、应用层三个层面顺序介绍了需求定位、业务梳理、建设内

容聚焦、建设路径分析与规划、实施步骤及运维策略制定等大数据项目建设过程中匹配大数据价值实现及落地应用的思维方式,本章下面的结构安排如下:12.2节介绍在进行大数据应用部署过程中,在数据方面应该具备的一些思维方式,包括数据全采样、数据交叉复用、数据云化存储;12.3节介绍在技术方面应该具备的一些思维方式,包括相关重于因果、效率优于精度、离线分析+实时应用;12.4节介绍在应用方面,即系统部署运维过程中应该具备的一些应用思维方式,包括数据质量溯源、服务和应用、开放和合作;12.5节对本章进行小结。

12.2 数据层

12.2.1 数据全采样

所谓数据全采样是相对于随机采样而言的。随机采样,也叫随机抽样,是指在采取子样时,对采样的部位或时间均不施加任何人为的意志,保证总体中每一个对象都有已知的、非零的概率被选入作为研究的对象,从而保证样本的代表性。利用随机采样进行数据研究是小数据时代的产物,具有坚实的数学基础,因为统计学证明随机采样的精确性来源于随机性。此外,随机样本得以流行(特别是在“小”数据时代)还有许多具体的实际问题使然:

- 1) 全体数据获取不可能或者成本太高。比如说你想知道北京有多少人吃过麦当劳,你不可能每个人都问一遍。
- 2) 资源限制:如果你要在极短时间内给出答案,即使不计成本你也做不到去问每个人。
- 3) 没有必要:因为我们可能只是需要知道吃过麦当劳的人数在北京总人口中的百分比,而且允许一定的误差,所以只要随机取样,通过统计分析就能够得到比较满意的答案。

根据统计学原理,如果我们随机取样1000人进行问卷调查,如果结果是某种百分比,那么得到的结果在置信度95%时的误差应该在3%左右。显然,在数据需要通过问卷调查的形式获取时,随机样本就成为我们的首选甚至是唯一的选择。

参见图12-1,抽样统计告诉我们的几个事实是:①样本量相同的情况下,置信水平越高,置信区间越宽。②置信水平相同的情况下,样本量越多,置信区间越窄。③置信区间不变的情况下,样本量越多,置信水平越高。也就意味着,样本量多到无穷的话,置信区间可以窄到0,置信水平高到1,数据全采样可以实现这样的追求,或许这也是大数据思维中包含数据全采样的原因之一。

数据全采样的缘由还有:①抽样采样的随机性很难保证。②只能从采样数据中得出事先

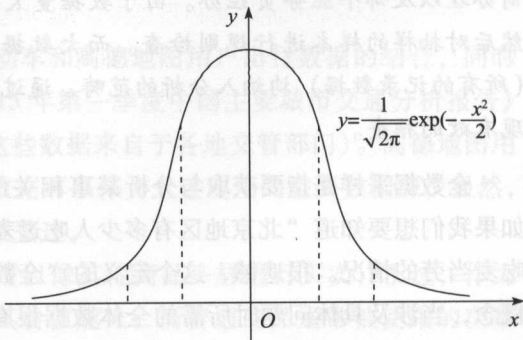


图12-1 抽样分析示意

设计好的问题的结果。③当人们想了解更深层次的细分领域的情况时，随机采样的方法就不可取了。

样本难以真正随机是因为你无法根据全体数据的分布去取样。比如你进行电话调查，有电话的人口就已经不是随机的样本空间；如果你要进一步知道海淀区30岁以下女性吃过麦当劳的人口比例，结果误差就会超过预期；问卷以外的内容你是无从知晓的。比如你突然想要知道吃过麦当劳的人中有多少同时吃过肯德基，但你却没有在调查中问这个问题，那你就根本得不到答案。

数据全采样是《大数据时代》一书中专门提出的概念，其原文是“抽样=全体”。

本示例是《大数据时代》中进行全数据论述时使用的一个案例：艾伯特·拉斯洛·巴拉巴西和他的同事想研究人与人之间的互动。于是他们调查了四个月内所有的移动通信记录——当然是匿名的，这些记录是一个为全美五分之一人口提供服务的无线运营商提供的。这是第一次在全社会层面用接近于“样本=总体”的数据资料进行网络分析。通过观察数百万人的所有通信记录，期望可以产生或许通过任何其他方式都无法产生的新观点。

事实上，这个在《大数据时代》中被认为是“第一次在全社会层面用接近于‘样本=总体’的数据资料进行网络分析”的“全数据”也就是一家移动公司四个月的通信记录。显然，四个月的数据怎么也谈不上是“全数据”。比较理性的解释应当是：《大数据时代》所说的“全数据”实际上就是我们通常所说的数据库数据，“全”体现在“包含了数据库中所有的记录”（而不是在这些记录中再进行抽样）。

仍然以运营商的应用场景为例，我国的任何一家运营商，其线下销售渠道有很多（有些属运营商自营，有些属于加盟），每个渠道均可以办理相关业务，出于很多原因，有些运营渠道会做一些违规业务，具体体现在用户办理业务时不按照既定的规则办理（比如办理了某个业务就不能办理另外一个业务，或者办理某个业务必须是一定级别的用户才可以），为了对违规业务进行监管，一般运营商都会设立专门的稽查部门进行渠道异常度监管。稽查部门每天面对的是所有用于办理业务的业务流水，这些流水记录着用户办理的业务类型、在哪个渠道商办理以及哪个业务员经办。由于数据量太大，根本无法逐一稽核，传统的做法就是抽样，然后对抽样的样本进行规则检查。而大数据思维应用在这个场景应该是把所有的业务流水（所有的记录数据）均纳入分析的范畴，通过相应的手段进行数据建模、异常预警等，借此实现有效的稽查。

全数据采样是指要获取与分析某事相关的所有数据，而不是依靠分析少量的数据样本。如果我们想要知道“北京地区有多少人吃过麦当劳”，这个全体数据就应该是北京地区所有人吃麦当劳的情况。很遗憾，这个定义的“全数据采样”是无法实现的。全体数据是个抽象的概念，当涉及具体问题时所需的全体数据很有可能并不存在。因此数据全采样思维重点是关注现有数据的“全”采样（而非抽样），而不是采样（获得）全体数据（甚至不是所有相关

的数据), 当然, 如果针对某个应用能够采集到更多、更全的数据进行后续的分析仍然是我们追求的。总结而言:

1) 随机样本与全数据采样并不是你死我活的绝对对立, 而是相互补充的, 即便有了全数据, 或者在一个既有的数据库里进行分析, 往往也有必要通过取样进行更有效的分析, 因为进行有寻优指标的抽样本身就是对面向目标应用的数据分析过程的优化。

2) 绝大多数所谓对全体数据的分析方法, 早在小数据时代就已经普遍存在, 并且随机抽样分析在大数据时代也还会继续展示其存在价值。

3) 绝大多数应用场景下, 所谓的全体数据可能都不存在, 在更多的情况下, 所谓的全体数据只是指企业的数据库数据全集, 不过在数据采集的时候, 应该有一种意识去引入更多的数据源, 以期获得更多的乃至更全的数据, 因为相对于既有的数据而言 (内部), 外部的世界或许更精彩。这也是在大数据时代, 时刻要抱有的“数据交叉复用”的思维。

12.2.2 数据交叉复用

说到底, 数据是指描述事物的符号记录, 而一个朴实的哲学原理是大千世界万事万物处于普遍的联系当中。例如, 记录某个事物 (记为 A) 的数据一定与某个相关事物 (记为 B) 具有某种关联, 这意味着对 B 事物进行分析的时候, 这个记录 A 事物的数据一定会被纳入数据采集的范围。由于 A 事物的关联物数量一般远超一个, 可能除了 B 之外, 还有 C 、 D 、 E 等, 所以, 当需要对 C 、 D 或者 E 进行分析的时候, 记录 A 事物的数据都会被纳入采集的范围, 反之亦然。因此, 事物本身的普遍联系决定了记录事物的数据的交叉复用的可能性和必然性。

从2014年起, 高德地图每个季度均会发布《中国主要城市交通分析报告》, 除了对中国的主要城市的堵塞状况进行排名外, 还把城市重大交通基建、交通政策和交通事件等对城市交通的影响纳入了关联分析范畴。截至2015年4月22日发布的《2015年第一季度中国主要城市交通分析报告》, 高德地图已经为中国45个城市进行了堵塞分析 (可以预计, 未来纳入分析的城市还会更多)。按照其官方说法: 高德将推出更多智能交通服务, 比如节假日出行预测、分析出行需求与限行政策的影响, 并量化不同政策治堵效果, 为交通管理部门提供决策辅助。

高德的交通出行数据来自于高德交通行业浮动车和高德地图用户出行数据的结合, 同时还结合各地交管部门的实时交通信息 (截至《2015年第一季度中国主要城市交通分析报告》的发布日, 实时交通信息支持全国114个城市, 这些数据来自于各地交管部门)。高德地图用户在享用高德地图免费服务的同时所产生的出行数据可以作为上述应用的重要数据源, 显然, 这些出行数据也可以用于其他的商业目标, 此处不赘述。

大数据时代的一个特征是, 数据就是资产, 可以像其他商品一样进行买卖, 不过, 作为一种商品, 数据的价值体现在被复用次数越多, 其价值越大; 而普通商品通常被使用的次数越多, 其价值就越小 (或许有一天成为古董, 会有价值的提升, 但原先的使用价值也可忽略

不计,更多的是它的收藏价值)。

大数据项目开展过程中经常被问,也会扪心自问的一些问题有:①有了一些数据,能做什么?②为了某个目标,还需要哪些数据?③有了这些数据,还能做些什么?

传统而言,做一个软件项目总是面向一个确定的目标需求的(做创新性产品除外,事实上,即便是做创新性产品也需要进行需求的挖掘和定位)。但是大数据这个概念的持续发酵,使得人们的思维发生了一个显著的变化,人们往往会首先有意识地自问:我有哪些数据,能做什么呢?为了便于讨论,我们将这些问题的应用场景设定为:有自营的平台或自有的一套业务系统,已经产生了一些数据。

针对这种应用场景,可以展开的工作至少有:

1) 对自身的价值定位和目标定位进行再梳理,梳理出对如下问题的解答:围绕既定的价值目标,目标用户是谁?按照公司的部署,拟响应哪些需求痛点?这些痛点的响应需要哪些相应的业务?各个业务如何响应以及各个业务有无对数据的需求等。类似的问题对手边没有数据,也没有应用目标的创新项目开展同样有效。

2) 对每一个既有的细分业务目标进行深挖,梳理出对如下问题的解答:服务对象是谁?相关的实体有哪些?现在如何提供服务的?有无已经成为痛点的需求?如果有更多的数据服务和应用服务是否可以更好地支撑既有的服务提供方式等。

3) 对既有业务进行重新梳理,尝试发现不同的业务部门之间有无重合的服务目标,梳理出对如下问题的解答:具有不同业务目标的业务(部门)是否有共性的数据需求?不同的业务(部门)是否存在相互依赖性?这种依赖性体现在哪些方面?是否可以通过数据服务方嫁接不同业务(部门)的有效协作等。

一旦梳理清楚,哪怕大致梳理清楚业务目标开始进行后续工作的时候,有必要再自问:为了这个目标,我既有的数据够吗?是否有更多的数据会更有助于目标的达成?需要哪些数据?数据源在哪里?如何获得?如果通过技术或者商务手段能够采集到更多的数据,自然有必要再迭代地问:有了这么多数据,有些数据还是大费周章获得的,是否可以服务更多的目标借此弥补额外数据采集的费用并彰显数据的价值?

12.2.3 数据云化存储

与传统的存储设备相比,云存储不仅仅是一个硬件,还是一个由网络设备、存储设备、服务器、应用软件、公用访问接口、接入网和客户端程序等多个部分组成的复杂系统。各部分以存储设备为核心,通过应用软件来对外提供数据存储和业务访问服务。

云存储系统的结构模型由4层组成:

(1) 存储层

云存储中的存储设备往往数量庞大且分布在不同地域,彼此之间通过广域网、互联网或者FC光纤通道网络连接在一起。存储设备之上是一个统一存储设备管理系统,可以实现存储设备的逻辑虚拟化管理、多链路冗余管理,以及硬件设备的状态监控和故障维护。

(2) 基础管理层

基础管理层通过集群、分布式文件系统和网格计算等技术,实现云存储中多个存储设备之间的协同工作,使多个存储设备可以对外提供同一种服务,并提供更大更强更好的数据访问性能。同时,通过各种数据备份以及容灾技术和措施可以保证云存储中的数据不会丢失,保证云存储自身的安全和稳定。

(3) 应用接口层

不同的云存储运营单位可以根据实际业务类型,开发不同的应用服务接口,提供不同的应用服务。比如视频监控应用平台、IPTV和视频点播应用平台、网络硬盘应用平台、远程数据备份应用平台。

(4) 访问层

任何一个授权用户都可以通过标准的公用应用接口来登录云存储系统,享受云存储服务。云存储运营单位不同,云存储提供的访问类型和访问手段也不同。

目前,业内企业针对云计算和云存储推出了很多种类的云服务,微软、EMC、亚马逊、谷歌等都是代表。作为一种理念,对每一个开展大数据项目的当事人而言,云存储是难以回避的一个技术选型,也是一种利益互补的部署实施模式。因为相比较于传统存储方法而言,云存储具有革新的意义和优势:

1) 高可靠性。云存储通过将文件复制并保存在不同的服务器上,很好地解决了潜在的由于硬件损坏而导致数据存取的不可用性问题。不论多好的硬件都可能出现故障,云存储知道文件存放的位置,在硬件发生损坏时,系统能自动将读写指令导向存放在另一台存储服务器上的文件,从而保持服务的继续性。

2) 容量扩展非常方便。云存储架构采用的是并行扩容方式,当容量不够时,只需采购新的存储服务器,容量即可增加,而且几乎没有上限控制。

3) 负载均衡。云存储能自动将工作任务均匀分配到不同的存储服务器上,从而可避免因个别存储服务器工作量过大而造成的性能瓶颈,这样可使整个存储系统发挥最大的功效。

4) 便于管理。对云存储管理者来说,即使再多的存储服务器也只是一台存储设备,管理人员只需在整体硬盘容量快用完时,增加采购存储服务器即可。而每台存储服务器的使用状况都可以很方便地在一个管理界面上看到。

在大数据部署中,选用云存储方案是一个流行及成熟的思路,不过针对具体的应用目标也要慎思云存储的可行性,因为相比较于上述的优点而言,云存储自有的一些瓶颈问题使得原本对此并不敏感的用户也会谈之色变。

1) 网络带宽的限制:真正的云存储系统是一个基于网络的、分布式的存储系统,只有宽带网络得到充足的发展,使用者才有可能享受到云存储服务。

2) 数据安全性:对于想要进行云存储的客户来说,安全性通常是首要的商业考虑和技术考虑,许多用户对云存储的安全要求甚至高于它们自己的架构所能提供的安全水平。但同时,许多大型的可信赖的云存储厂商所提供的云存储具有更少的安全漏洞,而且云存储所提供的

安全性水平往往要比用户自建的数据中心所能提供的安全水平还要高。

3) 应用存储的发展: 应用存储是一种在存储设备中集成了应用软件功能的存储设备, 它不仅具有数据存储功能, 还具有应用软件功能。应用存储技术的发展可以大量减少云存储中服务器的数量, 从而降低系统建设成本, 减少系统中由服务器造成的单点故障和性能瓶颈, 减少数据传输环节, 提高系统性能和效率, 保证整个系统的高效稳定运行。

12.3 分析层

对数据的分析是大数据系统的核心功能。随着数据量不断增加, 在有限的计算资源下, 我们无法对所有数据使用所有数据分析方法都分析一遍, 所以在分析的目标、技术、方法等方面, 都需要进行抉择, 而如何抉择就是分析思维方法使然。本节列举三种分析思维方法, 分别是相关重于因果、效率重于精度、离线分析 + 实时运行。

12.3.1 相关重于因果

因果关系和相关关系是说明事务之间的联系的两形式 (因果关系也是一种相关关系)。相关重于因果的观点出自《大数据时代》一书并很快被工业界和学术界普遍认可 (也存在一些争议)。

所谓相关性是指两个因素之间存在的联系。在统计学中, 两个随机变量 x 、 y (随机变量是对事件的数学抽象) 之间的相关关系用相关系数 ρ_{xy} 来表示:

- 1) 若 $\rho_{xy} = 0$, 则称 x 与 y 不相关。
- 2) 若 $\rho_{xy} \neq 0$, 则称 x 与 y 相关。
- 3) 当 $\rho_{xy} > 0$, 称 x 与 y 正相关, 如 $\rho_{xy} = 1$, 称 x 与 y 完全正相关 (备注: 未必是因果关系)。
- 4) 当 $\rho_{xy} < 0$ 时, 称 x 与 y 负相关, 如 $\rho_{xy} = -1$, 称 x 与 y 完全负相关。

所谓因果性是指原因 (cause) 和结果 (effect) 的必然联系。出于一般性, 作为 cause 的原因, 可能是一个事件或者一系列事实集; 而作为 effect 的结果, 可能是一个事件, 也可能是一个现象。通俗的理解就是平常所说的一因一果、多因一果、一因多果、多因多果、主要原因和次要原因等。为了便于后续的讨论, 因果关系描述为: $A \rightarrow B$, 因此, 因果性也是一种特殊的相关性, 关于因果性和相关性的哲学讨论参见 9.4.2 节, 本节重点考虑执行层面的具体事宜。

两个变量 A 和 B 具有相关性, 其原因有很多, 并非只有 $A \rightarrow B$ 或者 $B \rightarrow A$ 这样的因果关系。一个很常见的导致相关性的可能性是 A 和 B 都是同样的原因造成的: $C \rightarrow A$ 并且 $C \rightarrow B$, 那么 A 和 B 也会表现出明显的相关性, 但并不能说 $A \rightarrow B$ 或者 $B \rightarrow A$ 。数据挖掘领域经典的“啤酒—尿不湿”描述的啤酒和尿不湿是一种典型的相关关系, 而非因果关系; 大数据时代流行的“飓风来临前购买蛋挞的人一般也会同时购买手电筒”, 其中的蛋挞和手电筒是一个相关关系, 而“飓风和蛋挞”以及“飓风和手电筒”或许才是一个因果关系, 比如飓风来临, 不便于出

门采购,那么事先多备份一些蛋挞也是可以理解的,正如恶劣的飓风天气会导致普通照明的中断,准备一些手电筒也是可以理解的。

事实上,相关性的发现一直是应用一线的工作人员追求的目标,至于两者是否具有因果性,即使是一线工作人员,或许也并不关心。之所以在大数据时代,专门把因果性纳入与相关性的比较,可能的原因有:

1) 相关性本身就是从数据背后发现的知识和洞见,已经很有价值。对于商家而言,能够发现“啤酒—尿不湿”、“飓风—蛋挞”的相关性就已经能够为商家的商务决策提供辅助支撑,这或许就已足够。即使真的能发现某条因果性规律,或许随着时代的变迁,数据的广度和深度有了变化,原先的因果性也未必再适用了。

2) 如果能发掘出因果性,或许更有价值,不过因果性的发现所需要的成本太大,一般而言,因果性的发现除了需要在数据层面进行各种统计分析外,还需要配备其他的手段,比如因果关系调研,这是另外一门学问。

因果关系调研是指为了查明项目不同要素之间的关系,以及查明导致产生一定现象的原因所进行的调研。通过这种形式的调研,可以清楚外界因素的变化对项目进展的影响程度,以及项目决策变动与反应的灵敏性,具有一定程度的动态性。因果关系调研的目的是找出关联现象或变量之间的因果关系。其大致内容包括:①建立适当的因果次序或事件次序。②测量推测原因与结果间的相关性。③确认表面上合理的其他解释或原因性因素是否存在。调研问题的不确定性影响着调研项目的类型。在调研的早期阶段,当调研人员还不能肯定问题的性质时实施探索性调研,当调研人员意识到了问题但对有关情形缺乏完整的知识时,通常进行描述性调研(假设),因果性调研(测试假设)则要求严格地定义问题,寻找足够的证据来验证这一假设。

3) 相关性的发现能够为因果性的发现提供支撑。两个事务具有因果性的基础是这两个事务相关,这也是因果关系调研的准则之一。

在因果性调研中,一般对要解释的关系有种期望,如预期价格、包装、广告花费对销售额有影响。一个典型的因果性调研是改变一个自变量,然后观察因变量受到的影响,这种因和果的时序性是因果性调研的第一个准则。因果关系的第二个准则是存在相关关系,即两个考察的对象(事务)是按照某些可预知的方式一起变化,这是相关性发现的问题。也就是说,相关性的发现为因果性发现提供了技术铺垫。

需要注意的是,在某些方面,科学伦理不允许仅就数据谈论因果,否则数据歧视等一系列问题一定会产生,关于这一点,在3.4.3节有过介绍,此处不再赘述。

综上所述,关于因果性和相关性的应用提示至少有:

1) 因果性发现应当是人类追求自我解放的一个重要途径,但是因果性涉及的内容太多,甚至已经超出数据层本身。因此,在具体的大数据应用场景下,放弃“因果性”是一个不得

已的举措。当然,如果大数据的应用目标就是因果性发现,比如面向社会性问题的科学研究,那么因果性发现自然是必须考虑的,不过那是另外一个层面的事了。

2) 相关性发现是大数据分析中的一个重要内容,“相关重于因果”是一个准则,作为具体的一线人员应该有意识地发现数据层的相关性。传统的数据挖掘一般是基于人们自主的一个假设:属性 A 和属性 B 或许存在某种关联,然后利用专门的数据挖掘方法发现其中的关联性,借此达到对假设的验证,从而为后续的一些决策提供依据。在大数据分析场景下,或许我们要放弃传统的那种“白盒”的分析思路,而应该让分析更“黑盒化”,比如一个样本中有很多的属性,不妨把各个属性的两两关系都做一下关联分析,或许会发现有用或者有趣的一些规律。

2014 年夏季,阿里数据分析师在对阿里巴巴内衣销售数据分析后发现:购买大号内衣的女性往往更“败家”;65% B 罩杯的女性属于低消费顾客,而 C 罩杯及以上的顾客大多属于中等消费或高消费买家;新疆、香港和北京是购买 C 罩杯以上最多的地区,而黑龙江、浙江和江苏则是购买 A、B 罩杯最多的地区……

类似这样的规律看起来或许有趣,但分析其思路,其实都是根据销售数据中的各类数据项进行关联分析得出的结论。当然有些数据项(比如“败家”),是根据更多的消费数据标注的“标签”,是一个经过加工了的“数据项”而已。如何进行加工,则是数据建模的内容了。

12.3.2 效率重于精度

从数据分析的角度而言,精度一直是系统和算法追求的目标,为了这个目标,投入更多的精力进行算法的研究也是每一位数据工作者乐此不疲的事情。显然,对于任何一个算法,针对某一类场景都是有精度上限的。一般来说,越接近这个上限,提升精度的难度就越大。也有从另外的思路进行精度的提高,其中集成的思路是极为重要的一个里程碑。集成的思路大致可以描述为:用若干个不是那么高明的算法,通过某种策略上的集成,也能获得最终分析精度的提高。而且这个结论是有理论证明和事实检验的,并已成为工业界和学术界的共识。

盲人摸象的故事原型大致是:从前,有四个盲人很想知道大象是什么样子,可他们看不见,只好用手摸。胖盲人先摸到了大象的牙齿。他说:“我知道了,大象就像一个又大、又粗、又光滑的大萝卜。”高个子盲人摸到的是大象的耳朵。“不对,不对,大象明明是一把又薄、又大、又圆的大蒲扇嘛!”他大叫起来。“你们净瞎说,大象只是根大柱子。”原来矮个子盲人摸到了大象又长、又高的腿。而那位年老的盲人呢,却嘟囔:“唉,大象哪有那么大,它只不过是一根草绳。”原来他摸到的是大象又长又软的尾巴。四个盲人争吵不休,都说自己摸到的才是大象真正的样子。而实际上呢?他们一个也没有说对。

从数据分析的角度来看这个故事:

1) 每位盲人首先是一个“传感器”,感知了“大象”的(部分)特征(数据),比如故事

中的“又大、又粗、又光滑”、“又薄、又大、又圆”、“长长的、高高的”、“长长的、软软的”。这也符合“传感器”本身的特点，每一个传感器都是对每一类数据进行的数据感知（采集）。

2) 每位盲人又是一个“分类器”（决策），每位盲人根据自己采集的数据对“大象”进行一种分类，比如故事中的“大萝卜”、“大蒲扇”、“大柱子”、“绳子”。

3) 上面的故事原型中，盲人摸象的目标是想知道大象的样子，因此，如果能够把每个人感知到的特征融合在一起就是“又大、又粗、又光滑、又薄、又大、又圆、长长的、高高的、软软的”，显然这样的描述比任何一个人单纯感知到的内容要丰富得多，因此融合后的内容也丰富得多，在表现大象的样子这个目标上，利用更多数据源的数据，可以获得对大象更立体化的描述，这恰巧符合我们上述全数据采样的准则。

4) 上面的故事原型中，盲人摸象的目标是想知道大象的样子，因此，如果能够把每个人判断的中间结果融合在一起就是“大萝卜”、“大蒲扇”、“大柱子”、“绳子”，每一个判断结果都是每一种“算法”从各自不同的角度对“大象”这个对象进行的一个标签化。显然，这些标签化对于理解大象的样子是具有参考意义的，这也是大数据应用中，为对象实体进行标签化的一个原因。

5) 如果对上述的故事原型进行一些修改，事先不告诉盲人们将要摸的是什么动物或者物体，而是让盲人通过不同的采集方式判断这个动物或者物体。显然故事的其他情节会是类似的，从数据分析的角度出发，可以把每位盲人收集的数据整合在一起进行分类或者将每位盲人的判断结果融合在一起进行分析，前者对应着数据融合，后者对应着决策融合，都是一种集成思路。

在大数据应用环境下，数据分析方面对算法的追求已经不再是精度这一唯一目标，更重要的是这个算法如何应付“大数据量”及“数据处理大吞吐率”需求引发的对计算性能的追求。与计算性能相比，精度不是最关心的，因为利用更多的不是那么优异的算法，通过集成也能获得较好的结果。同时工业界和学术界普遍的共识是应用在大数据中的简单算法要比应用在小数据中的复杂算法要高明。

从1954年开始，人们便开始用计算机进行翻译工作；20世纪80年代后，IBM的研究员提出了新想法，不再教给计算机语言规则和单词，而是让计算机自己计算不同翻译可能结果的概率。于是这个名为Candide的项目将大约有300万句之多的加拿大议会资料译成了英语和法语并出版，利用这个“庞大的”语料库，让机器翻译能力提高了很多，但此后IBM公司再无此方面突破。而如今谷歌的这个语料库则是一个质的突破，后者使用庞大的数据库使得自然语言处理这一方向取得了飞跃式的发展。之后谷歌利用了互联网上获取的数十亿页、近千亿句的语料，再次让翻译质量飞跃式提高。从某种意义上讲，谷歌的语料库是一种退步，因为它的内容来自于未经过滤的网页内容，所以会包含一些不完整的句子、拼写错误、语法错误以及其他各种错误，也没有详细的人工纠错后的注解。但是，谷歌语料库是其他语料库的好几百万倍大，这样的优势完全压倒了缺点。

综上所述,关于效率优于精度的应用提示至少有:

1) 分析算法的高精度自然是应该追求的目标,不过在接近精度上限的精度提升的困难太大了,使用集成的手段,通过将不是那么强的算法集成在一起获得更高的精度是一个被认可的思路。

2) 大数据应用环境中,对数据分析的更高要求是性能高、效率高,借此能够实现对更大数据量、更大的数据处理吞吐量的需求响应,用更复杂的算法去追求更高的精度应该不是高明的策略,因为精度更高的算法往往其复杂度更高,或者是时间复杂度,或者是空间复杂度。而已经被认可的共识是大数据下的简单算法要优于小数据时代的复杂算法,因此大数据环境下,需要做的是如何让其更快。

3) 一般而言,大数据分析是“数据建模”+“模型应用”,前者是从大量的数据中,通过数据分析手法对数据进行建模,借此发现隐含在数据背后的知识和洞见;后者就是利用这个模型在具体的应用环境中进行面向目标的应用。因此,效率优于精度的准则更多需要数据建模侧的响应。这除了上述的原因要求数据建模必须使用更高效的算法外,还有一个重要的原因是:数据最终反映的是对象属性和物理行为,而任何一个实体对象或者物理环境都处在不停的变化当中,因此,利用“旧”的数据建模,然后利用这个模型去预测目标对象“未来”的输出,显然是有问题的。因此我们必须用更高效的方法完成对数据的建模,并且一直进行这样的迭代。唯有这样,才能使得建立的模型与最近的未来是相关的。可以想象,用两年前的数据建立的模型怎么能用在当前的应用场景呢?

4) 上述第3点的一个隐含提示是数据的活度要足够大,只有采集了活度足够大的数据,才能够保证数据建模的数据都是最新的,或者是与最近的未来的预测评估是相关的,否则即便有高效的数据建模方法也无法保证建立的模型能够为最近的未来进行有效的分析和评估,关于这点,会在下一节有更深入的介绍。

12.3.3 离线分析+实时运行

大数据的价值在于从数据中发掘知识和洞见,然后利用这些知识和洞见指导具体的生产实践,前者是数据建模(或知识发现)的过程,后者是模型应用的过程。相对而言,这两个过程的时间复杂度和空间复杂度远不对等,通常,建模的时间消耗和空间需求要远大于模型应用,因此,将数据建模与模型应用放在同一平台上有失公平。合理的做法是在离线情况下利用离线的数据进行数据建模,然后将模型提交给实时运行部分,实时运行部分利用最新构建的模型进行面向目标需求的应用。

最近有关在线训练的研究持续得到学界的关注,并成了一个研究热点,在线训练关注的是如何利用实时数据进行实时训练,其研究目标追求的是训练的实时性。本节关注的是当数据建模(学习过程)与模型应用(测试过程)在时间复杂度和空间复杂度不对等的情况下,采用离线分析+实时运行的策略进行数据建模和模型应用,追求的是训练与应用的解耦(彼此不相互影响),具体而言:

(1) 离线分析

离线分析的“离线”体现在离线分析的运行平台可以独立于项目部署实施的实时运行平台,另外,离线分析所依赖和需要的数据均是既有系统在某一个时刻运行的快照,离线分析所使用的数据都是这一个快照时刻之前的历史“数据”。

离线分析是指在后台(甚至与实施运维系统不在同一个计算平台)针对历史数据进行数据建模的过程。数据建模的结果评估方式有两种:一种是利用既有的数据集在进行建模的时候使用交叉验证方式,用部分既有(历史)数据建模,用另外一部分既有(历史)数据进行检验评估;另一种评估方法是将训练好的模型上线进行部署运行,然后通过实际运行的反馈数据评估数据模型。

交叉验证(Cross-validation)主要用于建模应用中,具体操作是:在给定的建模样本中,拿出大部分样本进行建模,留小部分样本用于对刚建立的模型进行检测,并求这小部分样本的预测误差,重复这个过程直至所有的样本都被预测了一次而且仅被预测一次。具体而言,常见的交叉验证形式有:Holdout验证、K折交叉验证和留一验证。

1) Holdout验证并非是一种交叉验证,因为数据并没有交叉使用。从最初的样本中随机地选出部分形成交叉验证数据,而剩余的就当作训练数据。一般来说,少于原本样本三分之一的数据被选做验证数据。

2) K折交叉验证(K-fold cross-validation)初始采样分割成K个子样本,一个单独的子样本被保留作为验证模型的数据,其他K-1个样本用来训练。交叉验证重复K次,每个子样本验证一次,平均K次的结果或者使用其他结合方式,最终得到一个单一估测。这个方法的优势在于,同时重复运用随机产生的子样本进行训练和验证,每次的结果验证一次,10折交叉验证是最常用的一种K折交叉验证。

3) 留一验证意指只使用原本样本中的一项来当作验证资料,而剩余的则留下来当作训练资料。这个步骤一直持续到每个样本都被当作一次验证资料。

需要注意的是,离线分析的数据都是某一个快照时刻前的历史数据,如果数据的活度很大,数据变化比较快,则意味着数据的建模过程是一个长期、持续和不断迭代的过程。

(2) 实时运行

实时运行的“实时”体现在系统需要和用户进行实时的人机交互,并根据人机交互目标提供实时的数据存取以实现用户的实时需求。一般而言,实时运行部分所依赖的模型都是在离线分析阶段构建好的,因此实时运行部分的计算负载往往要远小于离线分析环节。但正因为是实时运行,因此对人机交互的实时品质要求就很高,这意味着在计算架构的选型方面,一定要以能够提供快速、高效服务反馈为最终的目标。

综上所述,有关“离线分析+实时运行”的应用提示至少有:

1) 在应用层面,从逻辑上将大数据项目的部署实施划分为离线分析和实时运行两个部分,因此在体系架构的选型方面,应该针对离线分析和实时运行这两个不同的计算需求有针

对性地选择合适的计算架构。一般而言，离线分析更关注计算的性能，往往是无界面的；而实时运行部分往往需要考虑如何更有效地进行人机交互及前后台的通信（包括数据存取）。

2) 从逻辑上将大数据项目的部署实施划分为离线分析和实时运行两个部分，并不是说离线分析（数据建模）对效率不再追求，其要求甚至更高。特别是在数据变化比较快的场景下，数据变化快往往意味着数据所反映的规律也处于不断的变化中，针对此情况，唯有用最高效的方法快速地建立模型并不停地迭代建立模型然后将最新的模型提交给实时运行部分，才能保证所建立的模型是根据最新数据构建的。针对离线分析对计算性能诉求大的特点，计算模型的选择上要选取那些有助于高效分析的计算架构。

3) 针对实时运行部分，由于模型已经构建好，因此其计算的负载往往集中在与用户的交互以及与数据库的交互方面，这意味着在交互方式及实施架构的选择上要选取那些能够高效交互、高效存取的计算架构。以 Hadoop 和 Spark 对比为例，相对而言，基于内存交换的 Spark 在计算方面的性能要远优于基于磁盘交换的 Hadoop，因此在实时数据建模方面（对建模实时性有要求的场合），选择 Spark 要更加合适；而 Hadoop 这种分布式模型，虽然在计算能力方面远不如 Spark，但是其良好的分布式设计方案使得其非常适用于检索频繁的场合，这就意味着，当实时运行交互层面需要有大量的人机交互、数据存取需求的时候，相比较于 Spark，选择 Hadoop 更为合适。

4) 分布式计算是通过分而治之的策略将大的任务拆分成小的并行任务，从而达到高效计算的目的，流行的分布式计算架构包括：MapReduce、MapR、Spark 等。另外根据数据本身的特点，比如计算逻辑很简单，但数据量级比较大的场合，GPU 技术也是一个比较理性的技术选型，特别是当数据的结构化很明显时。当然现在也有将 GPU 和分布式架构融合在一起，即每个节点使用 GPU 技术，所有节点连接在一起做成一个分布式集成环境。

5) 特别需要强调的是，将大数据项目实施划分为离线分析和实时运行，仅仅是一种逻辑上的划分，实际部署实施中，两者是否使用同一平台是根据具体的情况具体分析。

12.4 应用层

12.4.1 数据质量溯源

大数据时代的一个共识是：数据是资产，数据是企业竞争力的源泉；数据是商品，数据是企业获益的一个重要手段。无论是从数据应用的角度还是从数据买卖的角度而言，数据质量（Data Quality）都尤为重要。关于数据质量的评估维度有：

1) 完整性：在许多场合也称为完备性，是指数据是充分的，任何有关的操作数据信息不存在缺失的状况，数据缺失的情况可能是整个数据记录缺失，也可能是数据中某个字段信息的记录缺失。不完整的数据所能借鉴的价值就会大大降低。显然数据的完整性是与前文提到的数据全采样有重叠的。

2) 一致性: 在许多场合也称为自治性。一致性是指数据遵循了统一的规范, 数据质量的一致性主要体现在数据记录的规范(数据编码的标准)和数据符合逻辑(多项数据间存在着固定的逻辑关系)上。

3) 准确性: 准确性是指数据记录的信息不存在异常或错误。和一致性不一样, 存在准确性问题的数据不仅仅只是规则上的不一致, 此外, 异常大或者小的数据也是不符合条件的数据。数据质量的准确性可能存在于个别记录, 也可能存在于整个数据集。

4) 实时性: 实时性是指数据从产生到可以查看的时间间隔, 也叫数据的延时时长, 但如果数据分析周期加上数据建立的时间过长, 就可能导致分析得出的结论失去了借鉴意义, 这也是前文提到大数据时代对数据的活度有所要求和追求的原因。

5) 真实性: 数据的真实性是指数据必须真实准确地反映实际发生的业务。IBM 认为真实性是当前企业亟待考虑的重要维度, 专门将真实性罗列为大数据的第5个特征(Veracity)。基于此认识, 他们投入精力在数据融合和先进的数学方法的研发上以进一步提升数据的质量, 从而创造更高价值。

6) 数据的使用质量: 数据的使用质量是指数据被正确地使用。这一点是与数据的分析耦合的, 任何一个高明的算法往往针对某一类数据、在某一个应用场景下是有效的, 如果采用了错误的或者不当的处理和分析手法, 即便数据的绝对质量很高, 也不可能得到正确的结果, 这也是对数据质量的监管要贯穿于数据的整个生命周期的原因。

7) 数据的存储质量: 数据的存储质量是指数据被安全地存储在合适的介质中。所谓安全是指采用了适当的技术手段以抵制潜在的攻击, 使数据免受破坏; 所谓存储在合适的介质是指数据的存取要便捷、高效。

8) 数据的传输质量: 数据的传输质量是指数据在传输过程中的效率和正确性。在现代信息社会中, 数据在异地之间的传输越来越多, 保证传输中的高效率 and 正确性至关重要。

上述的第1~4点特征关注的是数据本身的绝对质量, 第5~8点特征主要关注具体应用场景和应用过程中的数据质量, 其评估的依据是数据是否便于开发者使用以及数据是否能够与应用目标直接相关并带来价值。

2006年美国国会选举期间, 某政府工作志愿者在通过电话让已登记的选民来投票的过程中发现, 每十个选民中有三个是已经死去的人, 因此没有资格投票。显然, 针对“死去的人有没有选举权”问题本没有争议, 出现这样的问题也未必是“选民”恶意为之, 只是具有选民资格的数据与已死亡的人的数据没有进行逻辑检查才发生的。对于诸如保险公司、投资公司、基金公司、通信公司等拥有大量客户的服务类企业而言, 客户数据是其重要的财富来源。然而, 客户数据质量问题却一直是困扰企业开发新服务项目的绊脚石。有一项关于客户数据质量的调查研究发现: 平均而言, 8%~15%的客户数据记录存在各种问题, 例如各种证件号码输入错误、联系方式过期等, 其中有五分之一的数据问题是由于客户的死亡造成的, 其中一部分客户死亡时间超过十年却仍作为既有客户常年维护中, 显然这不仅是资源上的浪费, 也给潜在的“恶意违规”埋下了伏笔。

数据剖析和数据清洗是数据质量管理中的两个基本动作。数据剖析 (Data Profiling), 也称为数据考古 (Data Archeology), 是数据集内部为达一致性、单值性和逻辑性而进行的数值质量的统计分析及评估。数据剖析是 Olson 于 2003 年提出的概念, 其概念原型是使用分析技术来发现正确的、结构化的、有内容的、有质量的数据。数据清洗 (Data Cleaning) 是尝试通过移除空的数据行或重复的数据行、过滤数据行、聚集或转换数据值、分开多值单元等, 以半自动化的方式修复错误数据的过程 (对于是否过滤或修正一般要求客户确认)。

数据质量 (Data Quality) 是数据分析结论有效性和准确性的基础。无论是目前所在的大数据时代还是传统意义上的小数据分析, 对数据质量的追求如何过分都不为过。事实上, 也有专门进行数据质量管理和监管的商业化软件用于数据质量的保障, 一般而言, 这些商业化产品提供基于数据剖析、数据清洗 (过滤) 的交互数据转换工具。以下简单罗列几个典型的软件产品。

(1) DataCleaner

DataCleaner 是一个开源的数据质量分析工具, 用于管理和检测数据的质量。DataCleaner 包括一个独立的图形用户界面用于分析、比较和验证, 并监测 Web 应用。DataCleaner3.5 是一个主要的里程碑版本, 典型的特征有:

- 1) 支持连接到 Salesforce 和 SugarCRM。
- 2) 提升了向导和其他用户体验。
- 3) 支持作业的集群执行。
- 4) 新的数据可视化扩展和一个国家标识验证扩展。
- 5) 增加 Pentaho 数据集成作业调度和执行。

(2) OpenRefine

OpenRefine 是一款免费开源数据清洗工具, 最早源于众所周知的 Freebase Gridworks, 随后又演变为 Google Refine, 几年后又被社区接管, 在 2012 年 10 月变成了彻底开源的 OpenRefine。OpenRefine 是一款帮助用户转换数据集的工具, 它类似于传统 Excel 的表格处理软件, 但是工作方式更像是数据库, 以列和字段的方式工作, 而不是以单元格的方式工作。这意味着 OpenRefine 不仅适合对新的行数据进行编码, 而且功能极为强大。

(3) Wrangler

Wrangler 是一款由斯坦福大学的可视化组设计的用于清洗和重排数据的软件。它的动机是花更少的时间整理数据而将更多的时间花在学习数据中的知识。DataWrangler 是基于网络的服务, 使用非常方便。但是它的缺点是必须把数据上传到外部网站。也就是说, 对于敏感的内部数据, DataWrangler 并非合适的选择。

(4) DataStage

IBM WebSphereDataStage 为整个 ETL 过程提供了一个图形化的开发环境, 其主要功能包括:

- 1) 从范围最广的企业和外部数据源集成数据。

- 2) 合并数据验证规则。
- 3) 使用可伸缩的并行处理来处理并变换大量数据。
- 4) 处理非常复杂的变换。
- 5) 管理多个集成过程。
- 6) 可直接连接到作为源或目标的企业应用程序。
- 7) 利用元数据进行分析和维护。
- 8) 以批处理、实时或 Web service 方式操作。

(5) Informatica PowerCenter

Informatica PowerCenter 是 Informatica 公司开发的世界级的企业数据集成平台,也是业界领先的 ETL 工具。Informatica PowerCenter 使用户能够方便地从异构的已有系统和数据源中抽取数据,用来建立、部署、管理企业的数据仓库,从而帮助企业做出快速、正确的决策。此产品为满足企业级要求而设计,可以提供企业部门的数据和电子商务数据源之间的集成,如 XML、网站日志、关系型数据、主机和遗留系统数据源。Informatica PowerCenter 由服务器端组件和客户端组件组成。

服务端组件包括:

- 1) Informatica Service: PowerCenter 服务引擎。
- 2) Integration Service: 数据抽取、转换、装载服务引擎。
- 3) Repository Service: 知识库 Service, 管理 ETL 过程中产生的元数据, Repository 的数据存储在第三方数据库(如 Oracle)中。

客户端组件包括:

- 1) Administrator Console: 用于服务端各组件(Integration Service、Repository Service)的建立与维护。
- 2) Repository Manager: 知识库管理,包括安全管理等。
- 3) Designer: 设计开发环境,定义源及目标数据结构;设计转换规则,生成 ETL 映射。
- 4) Workflow Manager: 合理地实现复杂的 ETL 工作流,基于时间、事件的作业调度。
- 5) Workflow Monitor: 监控 Workflow 和 Session 运行情况,生成日志和报告。

综上所述,数据的质量溯源对应用的提示至少包括:

1) 每一个企业或多或少都存在垃圾数据方面的问题,这意味着所有的企业都应当加强重视,做好内部监控,严格执行例行的基本检查。从风险管理的观点来看,持之以恒地检查是最好的解决方案,利用一些商业评测软件或者自行开发相应的数据监控系统来自动识别不同系统的异常数据,并做好标记以方便检查,或许是一个相对务实的解决思路。特别需要注意的是:对于数据质量的监管贯穿于整个数据生命周期,是长期的过程。

2) 数据质量的监管应该对数据源、数据的获取以及数据的整个使用过程这三个方面进行监管,数据质量监管过程包括对数据(本身)质量监管和数据过程质量监管两个部分。数据质量监管的关键技术是数据语义的理解,设计一种数据规则提取方法,在此基础上实施数据

清洗和数据增强,保证数据本身的高质量,是一种静态数据质量监管方法。数据过程质量监管的关键是对历史操作的管理,包括对数据使用监管、对数据存储监管、对传输过程监管,是一种动态数据质量监管方法。通过静态质量监管和动态质量监管相结合,达到对数据正确性、准确性、一致性、时效性、冗余性的保障。

3) 数据本身的质量,或者说数据的绝对质量是受数据源约束的,这意味着针对数据源的遴选和评估是数据质量保障中的重要一环。特别是数据来自于互联网的时候,条件允许并符合分析目标要求的情况下,选择质量高的数据源往往会大大降低后续数据预处理的难度。而如果数据是从其他系统中 ETL 交换来的,就有必要在 ETL 过程中制定合适的数据规则引擎。大数据分析中有一种观点是数据交换应该变 ETL 为 ELT,其基本的思路是:将外部数据从其他系统导入到本系统后再进行目标应用驱动的转换。这种思路一方面兼具数据全采样的思路,另一方面也是一种数据入库以后再进行规则引擎评估和清洗的动机驱动。

4) 数据清洗 (Data Cleaning) 是数据整合过程中必须进行的一个复杂过程,通过检测和清除掉垃圾数据(包括不正确、过时、冗余以及不完整的数据)以保证数据的正确性、可靠性、完整性和一致性。数据清洗是保障数据质量的一个重要手段,不过应当注意的是:任何一种策略和技术手段的数据清洗必须以分析目标作为原始驱动。以冗余数据的剔除为例,冗余性本身也是反映物理世界的一个现实,因此如果分析目标与冗余性分析相关,就不能轻易地将冗余数据去除,即使去除,也要加上一些具有物理语义的标注,比如冗余度、数据源。因此比较保险的做法是把数据先保留下来作为原始库,然后进行相关的数据清洗、过滤、预处理等后续操作,然后对每一个涉及数据库的改变记录一个标签备注,以便在整个数据生存期内任何一个时间快照的数据均可回溯。

12.4.2 服务和应用

大数据本身就蕴含着服务的基因。

在大数据还没有得到媒体热捧的时候,云计算的概念就大行其道,其声势不弱于当前对大数据的热炒,当然大数据与云计算是天然的姊妹,这在本文中多个章节均有叙述,此处不再赘述。作为云计算的几个核心名词,比如软件即服务 (SaaS)、平台即服务 (PaaS)、基础设施即服务 (IaaS) 得到工业界和学术界的普遍认同。

事实上,在云计算还没有流行的时候,在 IT 公司中有一种企业生态是软件外包,其大意是作为软件外包的公司,为其他软件公司提供人力资源,承担其他公司专门的开发任务、测试任务等。曾几何时,甚至现在,软件外包形态仍然是政府支持的一种企业形态,从服务的角度来看,软件外包是一种典型的“人力即服务”模式。

在计算机科学与技术的研究领域,有一个方向叫作服务计算 (Service Computing),其本质是如何将计算以服务的形式封装、发布,如何根据具体的应用场景,自适应、自组织地选择和组合服务,为具体的应用场景提供服务。从某种意义上而言,这是一种典型的“计算即服务”的思想原型。

随着大数据时代的来临,不同专业背景的人,基于各自不同的技能素养一起形成了一个产业(研究)联盟,共同完成大数据的研究和落地实施。而服务的模式为不同专业背景的人之间的交流、合作提供了一个虚拟化的、透明的平台和界面。

于是在大数据环境下会发现各种商业模式的公司,比如专门进行数据收集的公司,以向其他公司提供数据服务作为企业的命脉;也有专门对数据加以标签化或者情报化,将这种信息提供给其他公司借此保证公司的盈利和运营,这是一种信息服务型公司;也有专门做各种算法研究的,然后将算法以服务的形式提供给第三方开发商,算是一种计算服务型公司(当然,在具体的部署过程中,或许是以插件、中间件等各种形式部署);即便是做具体大数据落地实施的公司,一般也会将数据服务和计算服务作为整个大数据管理平台的重要组成部分;至于后续的软件设施模式一般也遵循软件即服务、平台即服务、实施即服务的思路进行;甚至有专门做数据交易的服务平台,这种平台专门对各种数据源进行汇聚、价值评估、拍卖,建立买卖双方的第三方公证平台……任何一个大数据的落地应用都应该是多方资源合力的结果,各方提供各自能力范围的服务,通过服务的有效组合和协同共同实施大数据的落地。

综上所述,关于服务和应用的思维方式至少有:

1) 数据采集层需要有服务的思维:大数据部署实施的第一个环节是数据的采集与整合,从数据全采样的角度出发,应该采集更多的与目标应用有关的数据源数据,通过自主行为收集相关数据自然不可或缺,而同时,同数据服务商或者其他利益单位进行合作(采购、交换)也是快速获得数据的有效手段。如此,也可以将更多的精力用在后续的数据建模和业务应用的梳理和开发中。出于投资受限的约束,也可以考虑设计一个多边共赢的商业模式与相应的数据拥有者进行多边合作。

2) 前文提到的数据云存储思维也是一种将数据存储于管理层的服务思维。将数据存储以服务的方式外包给专门的机构,可以将重心放在系统的运维和商业价值的挖掘上,这对于自建存储系统而言,不仅降低了数据存储的成本开销,同时数据的安全也因为专业公司的投入而有所保障。甚至在整个系统的运维上,全部以云租赁的方式进行,将目标系统的运维挂接在专门的云服务上也是目前比较流行的一个做法。出于很多原因的考虑,在具体部署实施的时候,无法将所有的业务系统(或者数据)完全托管时,也可以通过混合云的部署实施方式,使得更“重”的计算负担和存储负担由专门的云服务提供商托管而将与自身业务更耦合的部分放在内部体系的私有云上,应该是一个比较切实的做法。

3) 在数据分析和业务梳理方面,应该有一种服务的意识。比较流行的做法是将数据和计算都以服务的方式封装,允许不同的业务部门,甚至不同的利益集团,比如第三方软件开发商,通过服务调用的方式获得数据或者获得对数据的加工、分析,获得更多的从数据中发掘知识和洞见的机会,从而获得更多的数据价值,这也便于拥有不同能力和资源的各方在大数据平台意义上获得更多的多边共赢,当然这需要商业模式的设计和梳理。

4) 任何一个大数据项目部署后,项目的运营团队必须跟进。这种跟进不仅仅是所谓的甲方以及乙方维护人员的跟进,往往还需要有专门的营运团队介入,比如技术保障、人力保障、

法务保障。这意味着，大数据项目部署后恰恰是多方力量合作的开始。

12.4.3 开放和合作

如前所述，大数据本身蕴含了服务的基因，潜台词是：开放与合作应该是大数据时代的不变主题。与前文所述的服务与应用的思维耦合的部分此处不再赘述，下文想从另外的角度讨论开放和合作的话题。

作为一个概念，大数据之所以能够获得几乎各界人士的普遍认同和关注的一个原因是：大数据能够产生价值，而大数据的价值有以下几个层面：

(1) 垂直应用价值：数据融合，价值细分

大数据环境下的需求特征呈现扁平化、垂直化和碎片化的趋势，在还没有足够资源去响应和匹配每一个细分需求的时候，“围绕垂直领域的细分需求进行数据的采集和整合，并在此基础上开发一套分析工具集和运维平台，为目标应用提供技术支撑”是大数据落地应用最可行、最务实的行动。即便是在“小”数据分析时代，围绕垂直应用的细分需求展开相关工作也是短期介入某个领域的最可行方案。事实上，越是在大数据时代，应用的垂直化需求以及针对垂直化需求的响应愈加明显。

以电商为例，垂直化的特征非常明显，垂直电商包括两类：一类是品类的垂直，如针对化妆品的聚美优品，针对图书销售的当当、亚马逊，针对生鲜市场的顺丰优选、天天果园，这类电商重在产业链整合，将标准品类做出特色，非标准品类做出品牌效应；另一类是目标人群的垂直，可通过挖掘特定人群的核心需求，进行品类扩张，满足用户综合购物的需求，如针对母婴儿童的宝宝树、红孩子，针对“屌丝”人群的凡客诚品。以搜索引擎为例，传统的泛化搜索有大家熟悉的 Google、Baidu、Bing，而在科研学术细分领域则有 Google 学术、知网空间，在旅游领域有携程旅行、去哪儿，在医疗卫生领域有丁香园、春雨医生，在金融财经领域有和讯网、新浪财经。这些在垂直行业中的佼佼者所提供的垂直搜索相比 Google、Baidu、Bing 的海量信息无序化，其特点是“专、精、深”，且更具有行业色彩而显得更加专注、具体和深入。

(2) 平台集成价值：信息分享，价值共赢

开展大数据项目，如果目标定位在垂直应用领域的细分市场，从目标规划上自然是没有错。但其获益或许仅仅是细分领域的细分市场的一部分，而如果能够利用既有的资源向直接相关的或者类似的领域或细分需求开拓，或者进一步说，如果能够将更多相关的垂直领域的细分需求融合为一个更大的平台，或许会为更大的获益埋下伏笔。

“ $1+1>2$ ”是超加环境下的一个经济准则，其基本含义是：在超加环境下，两个个体合作获益的总和要大于两个个体单独获益的累加和。仍然以电商为例，随着各家电商业务的不断渗透，垂直电商们的打法正悄然生变：从原来的追求“小而美”的垂直，到借助平台或推出开放平台逐步渗入到“高大全”的综合电商之路。以当今综合电商巨头京东为例，其就是

从垂直领域3C起家；美国电商巨头亚马逊也起家于细分品类图书。他们都是在相对细分的垂直领域积累了用户和口碑之后拓展品类，向着更高的目标进发。

(3) 生态协同价值：结构优化，价值发酵

任何一个经济角色都不是一个独立的个体，他一定是处于上下游产业链中的一环（一个节点），前面提到的垂直应用价值追求的是单个环节所能追求的价值，平台集成价值追求的是以这个环为中心的平行拓展所能追求的价值。而事实上，大数据更重要的价值凸显应该是以这个环节为中心的垂直拓展，即整个上下游产业链的无缝协同。其实前面提到的“重在产业链整合的垂直电商”就是一种借助产业链上下游整合的价值。

以热度不小于大数据的“工业4.0”为例，其追求的终极目标是物理信息系统融合，包括智能生产和智能工厂。智能工厂关注的是如何用更加智能、更加专业的生产设备和能力生产产品，其关注的是生产力；而智能生产关注的是产品设计、生产、仓储、物流、营销一体化的水平价值提升和生产过程涉及的上下游供应链的垂直价值提升。或许“工业4.0”追求的“价值网络的水平整合、端到端在工程上的数字整合以及网络化制造系统的垂直整合”能够切实带动整个社会产业链的联动，才是各国政府投注极大兴趣和关注度的缘由，至少是其中的一个重要原因。

上述三类价值层面事实上都隐式或显式地流露出开放和合作的必要性。

1) 在数据层的开放和合作。

数据是有价值的资产，是可以进行交换的商品。但如果一直处于封闭的状态，不加以应用或者仅限于本单位内部的应用，就很难最大化数据本身的价值。通过数据的开放和合作，可以在显著增强数据交叉复用的同时，进一步加强各边的黏性交互，创造出更多的商业价值。

新浪微博通过开放数据访问的API接口，允许第三方自由使用微博平台的数据。一方面进一步发挥和彰显了微博社交数据的价值，另一方面也进一步增加了第三方开发商及其服务对象对新浪微博数据的黏性和依赖，后者至少为更大可能的潜在合作起到了引流的作用，同时也因为这种开放和合作，许多创新性的产品不断被研发出来，从而进一步服务更多的人群。

阿里巴巴和新浪微博的合作不仅使阿里获得了一个重量级的广告平台，新浪微博所拥有的社交关系数据也弥补了阿里数据中的短板，使其获得了兴趣信息、关系信息等具有前瞻性价值的数据库。

电信运营商为摆脱“管道化”的困境，也纷纷在大数据方面挖掘价值潜力，比如法国电信除了利用大数据来提升本身的服务能力外，也向其他的领域扩张，通过和其他领域的合作，以寻求新的市场机会。

大数据时代进一步坐实了普通百姓对透明政府的需求，可以看到，多国政府都在进行以政务公开、信息开放为基础的透明政府建设，这除了表明了政府的自信外，也是民生建设、民心建设的重要保障。

2) 在数据分析层的开放和合作。

开源社区针对大数据的挑战和机遇开发的一系列开源产品，本身就是一种典型的开放精神使然，以开源精神驱动的开源系统和开源算法为第三方的开发极大地降低了开发成本，同时，开源社区的规模和品质也在不断扩张和提高之中。另一方面，将既有的算法以服务的方式提供，允许第三方自由地访问调用，并在此基础上进行二次开发，也为产品的快速迭代开发和有序演化提供了技术保障，前文提到的服务计算就是这种典型的思路和方法。

就数据分析技术研究层面而言，一个完整的数据分析流程包括数据预处理、特征提取和表示、数据建模与分析及交互展示。若干开源的标准数据集及算法平台为研究人员的算法研究提供了一个标准化的、可验证和评估的数据基础，为算法的研发提供了重要的基础保障。

Social Analysis (<http://socialysis.org/>) 就是中国十数所高校收集和整理的网络平台，为社会网络领域的研究者提供开放的数据集和算法，截至2015年5月，该平台已经积累了113个数据集和197种社会网络分析的相关算法。该平台的开放性还体现在它允许第三方研究者发布和提供共享的数据集和相关算法。这是社会网络研究者的基于开放和合作精神的一个举措，类似的例子还有很多，此处不一一赘述。

Weka 作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互式界面上的可视化。作为研究者或者数据分析工程师而言，如果想自己实现数据挖掘算法的话，可以参考 Weka 的接口文档，在 Weka 中可以很容易集成自己的算法甚至借鉴它的方法自己实现可视化工具。

Gephi 是一款开源的、免费的、跨平台的、基于 JVM 的复杂网络分析软件，主要用于各种网络和复杂系统。可用作探索性数据分析、链接分析、社交网络分析、生物网络分析等。同时 Gephi 也支持开发者编写自己感兴趣的插件，创建新的功能。

3) 在业务应用层的开放和合作。

任何一个软件产品或者业务系统往往只是迎合了具体领域的一个或几个细分需求，这就使得为了有效地响应一个完整的事务，往往需要若干个业务系统。在讲究数据交叉复用的大数据时代，即便是在同一个利益集团内部，不同的业务系统也是彼此独立的，每个业务系统的数据都被埋没在各个子系统中，呈现出一种新型的“数据孤岛”现象。或许这些项目的研发之初也会将数据的集成和融合纳入信息系统建设的目标，不过随着业务系统的推进，更多的精力被放在了业务的响应上而忽略了数据的融合，而当不同的业务系统（不同的业务系统往往是不同的开发商开发）上线运行以后，数据的融合就更加困难。

因此数据融合和集成是大数据时代不同业务系统有效集成的第一步，当然也是最关键的一步。可能的两种策略是：①在数据层融合。②在系统层集成。前者需要开放数据格式和访问权限，后者需要提供开放的数据接口（算是一种数据服务）。

国家科技支撑计划2015年度项目申报指南在“信息产业与现代服务业”中的“跨境电子商务服务技术研发与应用示范”中从产业结构优化和调整的角度提出了对业务协同、业务集

成的研究期待,其原文是:“研究提出跨境电子商务服务模式及解决方案,研发跨境电子商务基础信息可信保障技术、通关等基础业务协同技术、网络交易相关业务集成技术与平台并行示范应用,显著提升跨境网络交易基础信息可信度、通关等环节效率、网络交易相关业务集成度等。”

这仅仅是围绕跨境电商这一垂直领域的一个集成和协同需求,事实上在各行各业,任何一个利益集团内或许都存在这样的现象。

4) 在系统运维层的开放与合作。

大数据项目如何部署和运维是摆在每一个系统设计者面前的首要问题。系统的开发和部署自然需要不同部门、不同构建、不同业务系统之间的有效融合和集成,当然这些都是需要多边的开放和合作才能共同完成。此处需要强调的是,大数据项目部署实施上线后并不是大数据项目的结束,而是刚刚开始。因为只有有效地运营大数据项目(平台),才能真正获益。大数据项目(平台)的运维和运营需要多边力量的合力才能得以完成,至少需要的支撑团队包括:人力资源、法务资源、人才保障、基础设施保障、安全管理、商务支撑、咨询服务等。显然,如何有效组织这样的团队以便有效地支撑大数据项目的高效运行,这或许是一个与大数据技术本身无关的管理问题,其核心应该是合作和开放。

综上所述,关于开放与合作对于应用的提示至少有:

1) 合作的本质原因是一个人(一方)无法完成或者无法更好地完成某一个任务,这是产生合作意愿的最原始动机,而如何使这种合作持续、稳定地进行下去的重要保障是合作有获益以及获益分配各方均认可,前者需要在项目立项之初就有可行性研究和规划,后者需要相应的合作模式的设计来保障,这意味着在大数据环境下强调合作一定要把商业模式的梳理和设计从大数据项目一开始就纳入研究的范畴。

2) 开放是合作的基础,如果合作的各方不能如实地共享各边的资源 and 价值取向,本质上就难以获得全局收益的最终提高,也就失去合作的必要条件,这意味着在大数据项目开展之初,在商业模式的设计及合作模式的选取方面,必须选择一个激励兼容的合作机制。

3) 如果说大数据的价值源于垂直应用的价值、平台集成的价值、生态协同的价值这三个方面,应该说任何一个层面的价值获取,都需要有合作和开放的精神,而本身这三种价值的获取层面就代表了不同层面水平的或垂直的各个环节的合作和协同,当然是基于开放的精神。

4) 从数据的获取、数据的存取、数据的分析、应用的梳理到系统运维的各个环节,都需要有合作和开放的精神,即便是有些应用场景必须在某个封闭的环境下进行,也应该将这种合作和开放的精神在这个封闭的环境下彰显和体现。

12.5 本章小结

开展任何一个大数据项目,其建设步骤至少需要包括:需求定位、业务梳理、建设内容

聚焦、建设路径分析与规划、实施步骤及运维策略制定。在整个建设过程中，甚至部署实施后，必须要回答的问题清单至少包括：做什么？目标是否清晰？能做吗？是否多边可研？有什么？资源是否充足？如何做？配套能否落实？还能做什么？是否有价值？还需要什么？谁以及怎样运维？等等。仅回答上述问题或许还不够，在大数据建设的每一个阶段，都应该用大数据的方式去思考。

在数据层，我们应该进行数据全采样，尽可能多而广地去采集数据，但又不盲从于数据的“全”；我们应该充分进行数据交叉复用，进一步挖掘数据的价值，这事实上就耦合到业务目标的定位、业务梳理和运维中；我们应该充分考虑数据的云存储策略，在运维环境允许的情况下，数据云化存储一方面降低了大数据项目建设的基础设施的构建成本，另一方面，因为由专业的团队做专业的事情，也会进一步保障大数据项目的可靠落地。

在分析层，我们应该尽可能挖掘不同属性数据之间的相关性，其实在业务梳理的时候也应当充分考虑、梳理、调整不同业务之间的相关性，这在传统业务改造的时候尤为重要，显然，这也是与目标定位、业务和运维耦合的，同时我们应该用更快的计算效率弥补精确度不高的缺陷。已经达成共识的是：针对大数据的简单算法往往要高明得多。大数据的价值是发掘隐藏在数据背后的知识和洞见，然后用这种知识和洞见进行具体的实践，这意味着大数据的价值要依赖数据建模（知识发现）和知识应用两个环节，鉴于两者的计算复杂度有着天壤之别，在环境允许的情况下，通过离线建模进行知识发现和洞见挖掘，然后实时运用模型是一个有效的策略，但这会直接影响大数据系统的架构和实施方式。

在应用层，我们需要对数据的质量有着全生命周期的监管和维护意识，数据思维要体现在整个大数据项目建设和部署实施的各个环节中，更广一点，大数据项目的落地实施要有服务和合作的意识，本质的原因或许就是大数据项目的落地实施是多边利益团体共同协作的结果，大数据需要合作，需要彼此的服务，至于如何合作以及服务，则需要通过具体的商务或者商业模式来决定，这已不是本章的主题了。

对于传统企业而言，大数据项目的落地实施、部署应该从项目伊始，就布局和规划好整个大数据计划：

（1）建立数据计划

各个公司因业务模式的不同，需要涉及的数据也不同，是更关注产品还是企业运营抑或是人力的数据资源，这些问题需要在建立数据计划之初就做好考量。但涉及客户体验的数据，需要企业特别重视。或许当前这些数据还没有纳入业务体系的审核，但在传统企业比拼客户体验和服务意识的未来，这些数据经过挖掘和分析后将产生巨大的价值。

（2）建立数据管理和应用平台

企业做大数据，需要做好两个方面的基础。一方面是在IT基础设施上建立良好的数据处理结构，比如数据分布式存储、Hadoop。另一方面，企业要建立自己的数据管理和应用平台，包含数据的采集、数据库架构、分析模块、API接口等。需要注意的是，数据管理和应用平台的建立必须从公司业务出发，建设适合自己的平台。在数据中心建设方面，随着云计算和云

存储的出现,外部数据中心的成本已经大幅下降,数据存储的费用也不再是障碍,对于很多企业来说建立自己的数据中心并无必要。

(3) 建立数据团队

对于大型企业而言,自建数据挖掘的团队,无论是在成本控制还是业务响应机制上都相对有利。然而对于中小型企业来说,自建团队有时候并无必要,对这类型企业而言最重要的是将大数据思维融入企业的日常运营之中。

(4) 定制外部数据战略

有哪些外部数据会影响企业的业务发展?比如竞争品牌的售价、销售策略。这些都需要提前搜寻和沉淀。建立外部数据计划,企业可以通过公共渠道或者数据交换的方法来进行。如果是通过数据交换的方式来进行,这意味需要支付资金,关于这笔投资需要在一开始就制定相应的规划。设计一个好的商业模式,让各方在互赢的驱动下,降低前期的外部数据投资的直接现金流支出或许是一个好主意。

(5) 开放、再开放

企业保持开放共享的态度,不仅可以将自身存在的问题社会化,借助外部力量加以解决;另一方面,通过建立平等数据交换规则,还可以实现数据形式的共享。事实上,大数据的本质并非在于数据量的多少,而在于数据间的相关性。通过对整体数据流动的挖掘和分析,实现跨领域关联,才能最大化地彰显大数据的价值。

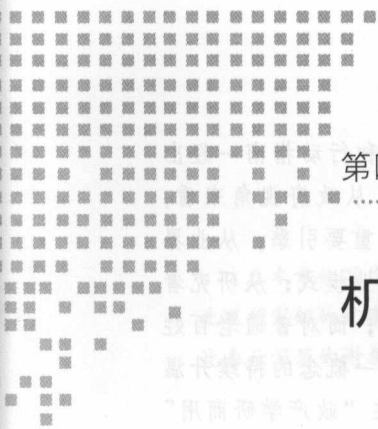
作为一种情操、修为和德行,“慎独”一词最早出自《礼记》。《礼记·中庸》如是说:“道也者,不可须臾离也;可离,非道也。是故君子戒慎乎其所不睹,恐惧乎其所不闻。莫见乎隐,莫显乎微,故君子慎其独也”;《礼记·大学》如是说:“所谓诚其意者,毋自欺也。如恶恶臭,如好好色,此之谓自谦。故君子必慎其独也”。

“慎独”原指不随波逐流之“格物、致知”以及不自欺欺人之“诚意、正心”的情操、素养和修为(解读维度很多,这是其中之一)。“格物、致知、诚意、正心”同样出自《礼记·大学》,指的是“历其事、求真知、意念诚实、心灵安静”,我国从日本引入“科学”这一词之前,将英文“Science”译为“格致”,其翻译思路应该就出自于此。

这对于大数据的应用提示或许在于:“科学的实践精神、务实的求知信念、不为浮华的本真、回归初心的审慎”是大数据分析师应该有的情怀。

本章参考文献

- [1] Kohavi R. A study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection [C]. Ijcai, 1995, 14(2): 1137-1145.
- [2] Olson J E. Data Quality: The Accuracy Dimension [M]. Burlington: Morgan Kaufmann, 2003.
- [3] 克托·迈尔·舍恩伯格,周涛. 大数据时代——生活、工作与思维的大变革 [J]. 人力资源管理, 2013(3): 136-136.



第四篇 Part 4

机遇及应用思索

.....第13章.....Chapter 13

大数据机遇

13.1 业引

从 20 世纪 90 年代开始,随着计算机技术的飞速发展,数据量呈指数级增长,“数据爆炸”已成为不争的事实。在“数据爆炸”的背后,隐藏着巨大的机遇。大数据的兴起,为各行各业提供了前所未有的机遇。在大数据的驱动下,企业可以更好地了解市场需求,优化生产流程,提高运营效率。同时,大数据也为科学研究提供了新的思路和方法,推动了人类对自然规律的探索。在大数据的浪潮中,企业和个人都面临着巨大的挑战,但同时也拥有着无限的机遇。只有抓住机遇,才能在激烈的竞争中脱颖而出。

1994 年,“互联网+”首次提出,标志着互联网开始与传统产业深度融合。2000 年,“互联网+”进入快速发展期,各行各业纷纷拥抱互联网。2001 年,全球互联网泡沫破灭,但“互联网+”并未因此停止脚步。2003 年,非典疫情爆发,加速了互联网的普及。2008 年,金融危机爆发,进一步推动了互联网的普及。2010 年,团购网站火爆全国,标志着“互联网+”进入爆发期。2011 年,微信诞生,开启了移动互联网时代。2013 年,余额宝横空出世,标志着互联网金融的兴起。2015 年,中国政府发布“互联网+”国家战略,为“互联网+”的发展提供了政策支持。

■ 第 13 章 大数据机遇

“互联网+”、“大数据”、“云计算”等概念层出不穷,成为人们热议的话题。在大数据的驱动下,企业可以更好地了解市场需求,优化生产流程,提高运营效率。同时,大数据也为科学研究提供了新的思路和方法,推动了人类对自然规律的探索。在大数据的浪潮中,企业和个人都面临着巨大的挑战,但同时也拥有着无限的机遇。只有抓住机遇,才能在激烈的竞争中脱颖而出。

2012 年伦敦奥运会开幕式开始之后,被誉为“互联网之父”的万维网发明者蒂姆·伯纳斯·李出现在场地中央。他在舞台上敲出一行字,此时大屏幕上显示“this is for everyone”,意喻万维网是送给世界上每一个人的礼物。

互联网、因特网、万维网这三个词语经常就有意无意地混淆,虽然只有太小的区别,但

2015年,“互联网+”这一凸显国家意志的战略规划和行动指南一经出台,便引发了“政产学研商用”各界的广泛关注和热议:从政府视角来看,“互联网+”是推动产业转型升级、提高创新力和生产力的重要引擎;从业界视角来看,“互联网+”是一种产业形态、应用模式或者商业模式;从研究者的视角来看,“互联网+”是一系列关键技术支撑下的应用;而对普通老百姓而言,“互联网+”意味着一系列的普惠。“互联网+”这一概念的持续升温势必会进一步助力和推动原本就炙手可热的大数据概念,在“政产学研商用”等多边的追捧、青睐下持续发酵。本篇尝试在对互联网的技术发展脉络及国际经济形势进行梳理的基础上,分析在“互联网+”概念被热炒及全民总动员的当代,大数据的潜在发展机遇和应用场景。本篇尝试通过对电子商务、工业4.0、互联网金融这三条主线的扼要描述和分析,厘清以下几个基本问题:“互联网+”的本质是什么?究竟是“互联网+X”还是“X+互联网”?“互联网+商务”“互联网+工业”“互联网+金融”的本质、门类及潜在机遇有哪些?作为一个数据分析师,在“互联网+”环境下应该有哪些情怀?

本篇包括一章内容:

第13章 大数据机遇 在对“互联网”及“互联网+”的相关背景及发展脉络进行简单梳理的基础上,本章尝试以大数据的视角理解和洞悉“互联网+”引发的机遇和挑战,并围绕电子商务、工业4.0、互联网金融这三条主线,给出一些大数据应用的场景示例和可能。

【关键字】 互联网+, 电子商务, 工业4.0, 互联网金融

大数据机遇

在本章的写作及润色过程中，得到了南京大学计算机科学与技术系及智能信息处理研究组的尹康、王强、徐鸣、李永春、陈鹏飞、李振兴等几位同学的协助，在此表示深深的谢意。

13.1 引言

1994 年，中国互联网时代开启；
1999 年，阿里巴巴创立，博客、QQ 诞生；
2000 年，百度成立，新浪、网易和搜狐在纳斯达克上市；
2001 年（前后），全球互联网泡沫；
2003 年，淘宝诞生；
2004 年，支付宝诞生；
2009 年，3G 牌照发放，全民微博时代开启；
2010 年，团购网站火爆全国；
2011 年，微信诞生；
2013 年，余额宝诞生，4G 牌照发放；
2015 年，中国政府发布“互联网+”国家战略。

毋庸置疑，最近这二十多年，中国人在互联网相关领域取得的成就举世瞩目，究其原因，后发优势使然，中国人的智慧使然，一个突飞猛进的时代使然。

2012 年伦敦奥运会开幕式开始之后，被誉为“互联网之父”的万维网发明者蒂姆·伯纳斯·李出现在场地中央，他在键盘上敲出一行字，此时大屏幕上显示“This is for everyone”，意喻万维网是送给世界上每一个人的礼物。

互联网、因特网、万维网这三个词语经常被有意无意地混淆，虽然没有太大的问题，但

是这三个词语所表示的内容是有区别的。这三者的关系可以简单地描述为：互联网包含因特网、因特网包含万维网。

互联网 (internet) 是指由若干设备相互连接而成的网络，互联网又可细分为广域网、城域网及局域网。

因特网 (Internet) 是互联网的一种，是从诞生于 1969 年的美国军用系统 ARPA 网逐步演化而来的，泛指“全世界”各国基于 TCP/IP 通信协定所建立的各种网络（该网络规模巨大，范围包括全世界而不单指某一地）。TCP/IP 是 Transmission Control Protocol/Internet Protocol 的简写，它定义了电子设备如何连入 Internet，以及数据如何在它们之间传输的标准，因而 TCP/IP 是 Internet 最基本的协议。协议采用了 4 层的层级结构，分别是：网络访问层、互联网层、传输层和应用层，每一层都呼叫它的下一层所提供的服务来完成自己的需求。

1) 网络访问层 (Link Layer) 在 TCP/IP 参考模型中并没有详细描述，只是指出主机必须使用某种协议与网络相连。

2) 互联网层 (Internet Layer) 使主机可以把分组发往任何网络，并使分组独立地传向目标，互联网层使用因特网协议 (Internet Protocol, IP)。

3) 传输层 (Transport Layer) 使源端和目的端机器上的对等实体可以进行会话。在这一层定义了两个端到端的协议：传输控制协议 (Transmission Control Protocol, TCP) 和用户数据报协议 (User Datagram Protocol, UDP)，前者是面向连接的协议，提供可靠的报文传输和对上层应用的连接服务；后者是面向无连接的不可靠传输的协议，主要用于不需要 TCP 的排序和流量控制等功能的应用程序。

4) 应用层 (Application Layer) 包含所有的高层协议，包括：TELNET (虚拟终端协议，TELEcommunications NETwork 的简称，允许一台机器上的用户登录到远程机器上，并进行工作)、FTP (文件传输协议，File Transfer Protocol 的简称，提供有效地将文件从一台机器传输到另一台机器上的方法)、SMTP (电子邮件传输协议，全称是 Simple Mail Transfer Protocol，用于电子邮件的收发)、DNS (域名服务，全称是 Domain Name Service，用于把主机名映射到网络地址)、NNTP (网上新闻传输协议，全称是 Net News Transfer Protocol，用于新闻的发布、检索和获取) 和超文本传送协议 HTTP (HyperText Transfer Protocol，用于在 WWW 上获取主页) 等。

只要应用层使用的是 HTTP 协议，就称为 WWW (World Wide Web，汉译为“万维网”，简称“WWW”)，因此万维网并不等同于因特网，而只是因特网所能提供的一项服务。万维网常简称为 Web，分为 Web 客户端和 Web 服务器程序，万维网可以让 Web 客户端（通常是用浏览器）访问 Web 服务器上的页面。

“互联网之父”是对在互联网的基础研究、应用推广等方面有杰出贡献的科学家的尊称。事实上，“互联网之父”不是一个人，而是一个群体，公认的有：温顿·瑟夫 (Vint Cerf，原名为 Vinton Gray Cerf)、罗伯特·卡恩 (Robert Elliot Kahn)、蒂姆·伯纳斯·李等。温顿·瑟夫和罗伯特·卡恩是 TCP/IP 协议的合作发明者，TCP/IP 定义了电子设备如何连入因特网，

以及数据如何在它们之间传输的标准，为互联网的发展奠定了革命性的基础。因为在互联网协议方面所取得的杰出成就，他们荣膺2004年的图灵奖，并于次年获得乔治·布什总统颁发的总统自由勋章，这是美国政府授予其公民的最高民事荣誉。蒂姆·伯纳斯·李的杰出贡献在于发明了万维网，他在1989年3月正式提出万维网的设想并于1990年12月25日，在日内瓦的欧洲粒子物理实验室里开发出了世界上第一个网页浏览器，开启了万维网时代。

互联网的拓扑结构如图13-1所示，可以理解为一些无生命的机器通过互联协议连接在一起，每个机器上有各自的数据或者信息，这些数据和信息或以文件、数据库的形式存放或者以网页的形式表示，显然后者是为了便于普通人的访问。

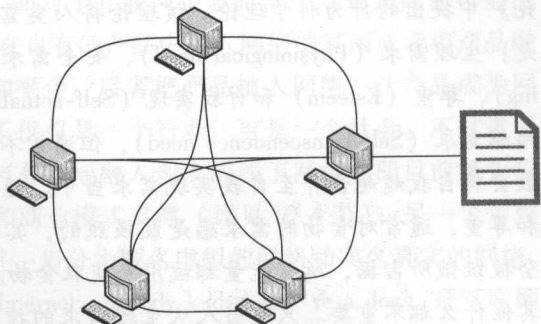


图13-1 普适互联网的节点

早期的互联网是信息的聚集地，也是信息共享的平台，这意味着，大部分用户是在网络上搜索并获得信息（当然也有一些人是专门从事信息的组织和推送的），因此当时的

互联网产品多是围绕信息的（有效）提供展开的，比如搜索引擎，无论百度、谷歌，其目的都是为了加速用户获取信息；再如新闻网站、网络广告，无非都是将传统的报纸、广告放在互联网这个工具平台上。

而随着技术的进步和时代的推进，上述的特征发生了明显的变化，比如：

1) 人与计算机的关系发生了变化。纵观人机交互的历史，本质上是人适应计算机到计算机适应人的历史。早期，计算机是人们用于科学计算或者数据管理的工具。然后计算机成为人的机能和功能的衍生，并因为互联网的迅猛发展，计算机成了（使用此计算机的）人与（使用另外一台计算机的）人的沟通媒介，也因为这个原因，社会性软件大行天下。所谓社会性软件是指构建于信息技术与互联网络之上的应用软件，在功能上能够反映和促进真实的社会关系的发展和交往活动的形成，使得人的活动与软件的功能融为一体。

社会性软件的核心内涵是应用模式从传统软件的“人-机对话”逐渐转变为社会性软件的“人-人对话”。传统软件主要是让机器完成文档处理或者信息获取，比如办公软件、ERP软件和浏览器，人们在使用这些软件（系统）的时候，感觉到软件（系统）是一种工具。社会性软件的主要功能是为了使网络中的人与人之间（而不是人与机器）进行对话。人们在使用该软件的过程中，感觉到的或者关注的是社会关系。

2) 人与互联网的关系发生了变化。人的社会性促进了社会性软件的发展。但同时，人的社会性，或者说人是发自内心被尊重、自我实现的内在需求决定了人更愿意在复杂的社会关系中构建自己的位置感。而互联网的扁平化特征进一步推动和促进了人的这方面的需求。以

典型的社会性软件博客（Blog）为例，博主书写博客，会不自觉地将自我体现在博客中，而读者在阅读博客的过程中，阅读的已经不是简单的日志信息，而是一个活生生的人（博主）。再如，社交网络平台上，人们会毫无顾忌地填写自己的生辰八字、教育履历、工作单位甚至随时随地分享自己的身边事及自己的位置、感悟等。上述实例表明人与互联网的关系从最早期的信息受众成为一个主动的信息提供者，从而成为互联网的一部分。

马斯洛需求层次理论是由美国心理学家亚伯拉罕·马斯洛于1943年在论文《人类激励理论》中提出的行为科学理论。该理论将人类需求像阶梯一样从低到高按层次分为五种，分别是：生理需求（Physiological need）、安全需求（Safety need）、爱和归属感（Love and belonging）、尊重（Esteem）和自我实现（Self-actualization）五类，在自我实现需求之后，还有自我超越需求（Self-transcendence need），但通常不作为马斯洛需求层次理论中必要的层次，大多数会将自我超越合并至自我实现需求当中。通俗理解：假如一个人同时缺乏食物、安全、爱和尊重，通常对食物的需求感是最强烈的，其他需求则显得不那么重要。此时人的意识几乎全被饥饿所占据，所有能量都被用来获取食物。在这种极端情况下，人生的全部意义就是吃，其他什么都不重要。只有当人从生理需求的控制下解放出来时，才可能出现更高级的、社会化程度更高的需求。

互联网的特征可以用三个字来表示，分别是“互”“联”“网”，简单介绍如下：

特征一：“互”

互联网的“互”指的是互联网具有交互性，这种交互不仅是指计算机与计算机之间的通信，更是指使用计算机的人与处于网络中的计算机之间的交互，比如互联网早期产品BBS以及随着技术推进而出现的微博、微信都显示了人与机、人与人之间的交互性。

特征二：“联”

互联网的“联”指的是互联网具有连接性，互联网的基础是节点与节点之间具有连接性，最早期节点与节点之间是通过有线连接在一起的。随着无线通信技术的发展，节点之间的连接可以通过无线连接。无论是有线连接还是无线连接，节点往往都是计算机，目标都是通过计算机与计算机的相连实现人机交互或人人交互。而作为互联网的延伸和扩展，基于智能感知、识别技术与普适计算等通信感知技术得以迅猛发展的物联网（Internet of things）关注的是物与物的连接，即用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信，实现的目标是人与物相连、物与物相连。

特征三：“网”

互联网的“网”指的是互联网具有网络性，即互联网是一个由一组节点集合和节点与节点之间的边集合组成的（复杂）网络。其中，节点的类型可以是计算机，也可以是有通信能力的“物”、人等；边的类型可以是无向的，也可以是有向的，后者还可以进一步分为单向的，或双向的。

互联网的上述特点使得互联网（相关）行业具有典型的媒体属性、社会属性、技术属性

和商业属性。

(1) 媒体属性

媒体是指实现信息从信息源传递到受众者的工具、渠道、载体、中介物或技术手段,包含两层含义:一是承载信息的载体;二是指存储、呈现、处理、传递信息的实体。互联网不是媒体,但是互联网的一些内容具有典型的媒体属性特征,如网站、微博(大V)、微信(公众号)。

(2) 社会属性

如前所述,互联网本质上是一个复杂网络,而对于互联网企业而言(特别是面向行业的互联网企业),有效拥抱互联网的策略至少包括在自有的产品中植入网络的基因或者将产品融入网络中,前者将互联网作为产品的延展工具和平台,后者把产品植入网络,让产品成为网络的一部分。因此,有一种说法是互联网行业不仅仅是一个行业,更是一个社会。不过需要注意的是,互联网是一个虚拟的线上网络,如何有效地融入现实的线下网络,即目前极为流行的O2O,还需要专门的技术手段、应用模式和商业模式支撑(详见10.4节);另一个需要注意的是:即便在虚拟的线上网络中,也可以进一步分为匿名虚拟的网络和实名真实的网络。“在互联网上,没人知道你是一条狗”(On the Internet, nobody knows you're a dog)是互联网上一个极具趣味性的语句,描述的就是典型的匿名虚拟社交,而近些年以熟人社交为主的微信则将线下的真实关系搬到互联网中,并取得了极大的成功,此处不再赘述。当然,“匿名/虚拟”与“实名/真实”这两个不同的互联网世界存在着界限,也存在着交集,也有很多的商家出于商业价值的追求,专门尝试在这两个网络中做相关工作。

(3) 技术属性

毋庸置疑,互联网是在一系列的技术积淀基础上,逐步发展起来的、连接人与人之间沟通交互的工具和平台。因此,对于任何一个互联网企业而言,如何在植入互联网基因的前提下,满足用户的需求是企业立于不败之地的根本。这自然涉及一系列相关技术理论(此处不赘述)。因此,任何一个互联网企业都可以看成是以技术驱动的企业。

(4) 商业属性

显然,对于以盈利为目的的商家而言,互联网提供了彼此信息对称透明、沟通简单扁平的便利。如何挖掘互联网带来的“红利”是商家关注的重点,比如如何找到客户、如何对客户进行立体画像、客户需要什么、客户偏好是什么,所有这些问题是技术话题,也是商家从互联网中获益的重要手段。

我们生活在互联网时代,互联网将继续改变和改善我们的生活。每一个生活在这个时代的自然人、企业法人都应该努力地拥抱这个时代。问题是:我们拥抱得足够吗?怎样去拥抱?如何拥抱得更好?2015年3月15日,中国政府发布“互联网+”国家战略,从国家意志层面面向全国发起了总动员。

本章在对“互联网”及“互联网+”的相关背景及发展脉络进行简单梳理的基础上,尝试以大数据的视角理解和洞悉“互联网+”引发的机遇和挑战,并围绕电子商务、工业4.0、互联网金融这三条主线,给出了一些大数据应用的场景示例和可能,本章下面的结构安排如

下：13.2 节简单介绍“互联网+”涉及的相关概念；13.3 节简单介绍“互联网+商务”相关内容，重点介绍电子商务、移动电子商务及跨境电子商务并给出了具体的应用提示；13.4 节简单介绍“互联网+工业”相关内容，重点介绍工业4.0引发的各个方面的嬗变，并给出具体的应用提示；13.5 节简单介绍“互联网+金融”相关内容，重点介绍面向投资、融资、支付及其他类型的互联网金融，并给出具体的应用提示；13.6 节对本章进行小结。

13.2 互联网+

2015年3月5日，李克强总理在两会政府工作报告中，提出了“互联网+”的战略规划，相关上下文是：制定“互联网+”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展。

这段文字的几个关键字是：互联网+、移动互联网、云计算、大数据、物联网、电子商务、工业互联网、互联网金融。其中：

1) 移动互联网、物联网都是泛在的互联网（或者说是互联网的不同基础支撑），前者重点在于互联网连接的“无线”，由移动通信技术加以保障。后者重点在“物—物相连”，由具有传感和通信功能的智能终端加以保障。

2) 云计算、大数据均属于技术支撑的范畴，也是本书的关注重点。

3) 电子商务、工业互联网和互联网金融应该是中国政府（在提出“互联网+”时）拟定的重点和优先扶持发展的行业。

腾讯马化腾是“互联网+”的积极提倡者和实践者。早在2013年，他与马云、马明哲在上海一起推出众安保险时，就提及“互联网+”的概念。在2015年5月22日，马化腾主持编著的《互联网+国家战略行动路线图》发行，该书以腾讯官方视角，专门解读李克强总理“互联网+”国家战略。他认为：所谓“互联网+”，就是以互联网为基础，利用信息通信与技术和各行业的跨界融合，推动产业转型升级，并不断创造出新产品、新业务与新模式，构建连接一切的新生态。

阿里马云认为“互联网+”是以互联网为主的一整套信息技术（包括移动互联网、云计算、大数据技术等）在经济、社会生活各部门的扩散应用过程。

百度李彦宏认为“互联网+”是互联网和其他传统产业的一种结合模式，比如O2O。

小米雷军认为“互联网+”是用互联网的技术手段和互联网的思维与实体经济相结合，促进实体经济转型、增值、提效。

事实上，不同公司对“互联网+”都有自己价值观驱动的理解和定义，但趋同的观点是：“互联网+”是一种产业形态、一种应用模式或者一种商业模式，其实施路径是将互联网基因植入不同行业，或者对某个行业进行革命性的颠覆，或者对某个行业的产品进行颠覆、优化、改善，或者对行业中的业务流程进行优化、改善。

一般用“互联网+X”来具体地称谓“互联网+”，此处的“X”指的是行业，比如互联网

金融、互联网医疗、互联网社区、互联网交通、互联网电视、互联网餐饮。当然也有观点说不应该是“互联网+X”，而应当是“X+互联网”，本文暂时不加以区别，后文会有具体的分析。

行业和产业是两个容易混淆的概念，以下进行简单介绍：

所谓行业指的是以生产要素组合为特征的各类经济活动，行业是根据生产力三要素（劳动者、劳动对象、劳动资料）不同排列组合的各类经济活动的特点划分的；国家统计局制定的国民经济行业分类（GB/T 4754—2011）将行业分为农、林、牧、渔业，采矿业，制造业，电力、热力、燃气及水生产和供应业，建筑业，批发和零售业，交通运输、仓储和邮政业，住宿和餐饮业，信息传输、软件和信息技术服务业，金融业，房地产业，租赁和商务服务业，科学研究和技术服务业，水利、环境和公共设施管理业，居民服务、修理和其他服务业，教育，卫生和社会工作，文化、体育和娱乐业，公共管理、社会保障和社会组织，国际组织，共计20个行业（详见官网 <http://www.stats.gov.cn>）。

所谓产业指的是各类行业在社会生产力布局中发挥不同作用的称谓，这是按照生产力布局的宏观领域来进行划分的。目前在国际普遍流行的是三次产业划分思路，即按照人类生产发展的历史顺序：第一农业、第二加工制造业、第三服务业来划分，并用来反映国民经济中各类活动的不同特征。第一产业专指提供生产资料的产业，泛称农业，包括农业、林业、畜牧业、渔业；第二产业专指加工产业，泛称工业，包括采矿业、制造业、水电气生产及供应业、建筑业；第三产业专指非物质生产部门，泛称服务业，包括交通、仓储和邮政业、通信、计算机和软件业、批发和零售业、住宿和餐饮业、金融业、房地产业、租赁和商务服务业、科学研究、技术服务和地质勘查业、水利、环境和公共设施管理业、居民服务和其他服务业、教育、卫生、社会保障和社会福利业、文化、体育和娱乐业、公共管理和社会组织、国际组织。

理论上而言，“互联网+X”中的“X”可以为上述产业中的任何一个行业，但需要注意的是：

1) 不同行业的固有特点会导致在不同的行业进行“互联网+”的实施时所面临的挑战和困难也不一样，这意味着在不同行业进行“互联网+”需要不同的实施策略和能力的同时，实施代价也不一样，甚至可以说，并不是所有的行业都适合“互联网+”。

2) 与“互联网+X”并行的另外一个说法叫作“X+互联网”，两种说法的本质都是通过互联网基因与X行业有效融合，提高产品的生命力、竞争力并最终达到获得最大收益的目的。两种称呼的本质区别是在于以谁（互联网或者X）为核心进行融合改造，通常在行业X更核心的情况下，人们愿意用“X+互联网”来称呼，比如农业互联网、工业互联网（相对而言，分属于第一、第二产业的农业和工业，其互联网改造的难度要远大于作为第三产业的服务业）；而对于互联网改造难度相对小的行业，人们更愿意用“互联网+X”来称呼，比如互联网交通、互联网金融、互联网餐饮。

3) 即便是针对同一行业，在进行“互联网+”的实施时，也会因为实施策略、思路及市场环境等约束条件的不同而在实施成效上有本质的差别。

2015年7月1日，国务院公开发布《国务院关于积极推进“互联网+”行动的指导意见》（国发〔2015〕40号），从国家层面明确了行动要求、重点行动、保障支持，该建议明确地提出下一步重点行动集中于创业创新、协调制造、现代农业、智慧能源、普惠金融、益民服务、高效物流、电子商务、便捷交通、绿色生态、人工智能共11个方面，进行有针对性的“互联网+”实践和示范，如表13-1所示。

表 13-1 国发〔2015〕40号建议重点行动

序号	互联网+	备注
1	创业创新	推动各类要素资源聚集、开放和共享，大力发展众创空间、开放式创新等，引导和推动全社会形成大众创业、万众创新的浓厚氛围，打造经济发展新引擎
2	协调制造	推动互联网与制造业融合，提升制造业数字化、网络化、智能化水平，加强产业链协作，发展基于互联网的协同制造模式。在重点领域推进智能制造、大规模个性化定制、网络化协同制造和服务性制造，打造一批网络化协同制造公共服务平台，加快形成制造业网络化产业生态体系
3	现代农业	利用互联网提升农业生产、经营、管理和服务水平，培育一批网络化、智能化、精细化的现代生态农业新模式，形成示范带动效应，加快完善新型农业生产经营体系，培训多样化农业互联网管理服务模式，逐步建立农副产品、农资质量安全追溯体系，促进农业现代化水平明显提升
4	智慧能源	通过互联网促进能源系统扁平化，推进能源生产与消费模式革命，提高能源利用效率，推动节能减排。加强分布式能源网络建设，提高可再生能源占比，促进能源利用结构优化。加快发电设施、用电设施和电网智能化改造，提高电力系统的安全性、稳定性和可靠性
5	普惠金融	促进互联网金融健康发展，全面提升互联网金融服务能力和普惠水平，鼓励互联网与银行、证券、保险、基金的融合创新，为大众提供丰富、安全、便捷的金融产品和服务，更好满足不同层次实体经济的投融资需求，培育一批具有行业影响力的互联网金融创新企业
6	益民服务	充分利用互联网的高效、便捷优势，提高资源利用效率，降低服务消费成本。大力发展以互联网为载体、线上线下互动的新兴消费，加快发展基于互联网的医疗、健康、养老、教育、旅游、社保等新兴服务，创新政府服务模式，提升政府科学决策能力和管理水平
7	高效物流	加快建设跨行业、跨区域的物流信息服务平台，提高物流供需信息对接和实用效率。鼓励大数据、云计算在物流领域的应用，建设智能仓储体系，优化物流运作流程，提升物流仓储的自动化、智能化水平和运转效率，降低物流成本
8	电子商务	巩固和增强我国电子商务发展领先优势，大力发展农村电商、行业电商和跨境电商，进一步扩大电子商务发展空间。电子商务与其他产业融合不断深化，网络化生产、流通、消费更加普及，标准规范、公共服务等支撑环境基本完善
9	便捷交通	加快互联网与交通运输领域的深度融合，通过基础设施、运输工具、运行信息等互联网化，推进基于互联网平台的便捷化交通运输服务发展，显著提高交通运输资源利用效率和管理精细化水平，全面提升交通运输行业服务品质和科学治理能力
10	绿色生态	推动互联网与生态文明建设深度融合，完善污染物检测及信息发布系统，形成覆盖主要生态要素的资源环境承载力动态检测网络，实现生态环境数据互联互通和开放共享。充分发挥互联网在逆向物流回收体系中的平台作用，促进再生资源交易利用便捷化、互动化、透明化促进生产生活方式绿色化
11	人工智能	依托互联网平台提供人工智能公共创新服务，加快人工智能核心技术突破，促进人工智能在智能家居、智能终端、智能汽车、机器人等领域的推广应用，培育若干引领全球人工智能发展的骨干企业和创新团队，形成创新活跃、开放合作、协同发展的产业生态

由表13-1，建议书中指明的重点行动涉及的11个行业，属于第一产业的是现代农业，属于第二产业的是协同制造，属于典型的第三产业的行业包括：智慧能源、普惠金融、益民服

务、高效物流、电子商务、便捷交通、绿色生态。

创业创新和人工智能是两个比较特殊的行动方向,前者是目前我国鼓励和支持的国家战略(全称是“全民创业、万众创新”),后者是计算机科学与技术的一个研究方向和关键技术。重点行动的本意是希望通过互联网加速人工智能的研究和产业化推动,提高精密制造的能力(工业机器人),并使相关领域的产品,特别是能带领上下游产业链联动的产品(智能家居、智能终端、智能汽车)更具智能性。本书在不同的章节都有对人工智能的介绍,下文聚焦讨论“大众创业、万众创新”。

“大众创业、万众创新”这一关键词最早是李克强总理在2014年9月的夏季达沃斯论坛上提出的,相关上下文是:……要在960万平方公里土地上掀起“大众创业、草根创业”的新浪潮,形成“万众创新、人人创新”的新态势。此后,他在首届世界互联网大会、国务院常务会议和各种场合中频频阐释这一关键词。

李克强总理在2015年政府工作报告中又提出了“大众创业、万众创新”,在政府工作报告中如此表述:“推动大众创业、万众创新,既可以扩大就业、增加居民收入,又有利于促进社会纵向流动和公平正义”。在论及创业创新文化时,强调“以简政放权的改革为市场主体释放更大空间,让人们在创造财富的过程中,更好地实现精神追求和自身价值”。

由政府层面倡议和推动、全国上下积极响应并迅速发酵“全民创业、万众创新”,理性的原因或许是:

1) 缓解就业率问题。这里的就业率的缓解包括两个方面的内容:一方面现有社会结构下既有的劳动力分配不均(不合理问题),通过自行创业或者加盟别人创新创业的公司,可以缓解和改善既有的劳动力的有序分配;另一方面,成为发展趋势的“互联网+”如果在各个行业实施成功,一定会带来行业的革命性颠覆或者(仅仅是)完善,但劳动力的分流是必然(有些岗位因为“互联网+”消失或者劳动力需求量降低),这意味着“互联网+”引起的剩余劳动力必须有更多的出口加以消化,创新创业无疑是一个正能量的策略。

2) 提升幸福感。按照前文提到的马斯洛需求层次理论,人都是有被认同和自我实现的需求的,因此通过创新创业获得的被尊重和被认同会引发发自内心的幸福感的提升,而这显然是中国梦在老百姓层次的实现路径之一。

3) 倒逼体制改革。毋庸置疑,中国现有的多边体制需要进行结构化调整,但是在进行体制改革和完善的过程中,如果没有实时的反馈,势必会引发改革成本的提高,因此通过制定“大众创业、万众创新”目标,在这个创新创业大环境下,创业期间的相关行为倒逼各种体制结构化调整,最后达到多边体制的完善和优化。

13.3 电子商务

13.3.1 电子商务概述

电子商务通常是指在全球范围的商业贸易活动中,在Internet环境下,买卖双方不谋面地

进行各种商贸活动,实现消费者的网上购物、商户之间的网上交易和在线电子支付以及各种商务活动、交易活动、金融活动和相关的综合服务活动的一种新型的商业运营模式。事实上,由于各国政府、学者、企业界人士根据自己所处的地位和对电子商务参与的角度和程度的不同,他们对于电子商务的定义和理解也不完全相同。比较公认的一种观点是:电子商务=商流+物流+资金流+信息流,其中:

1) 商流:指的是物品所有权转移的过程,是物流、资金流和信息流的起点和前提,没有商流一般不会发生物流、资金流和信息流。

2) 物流:是物品从供应地向接收地的实体流动过程中,根据实际需要,将运输、仓储、采购、装卸、包装、流通、配送、信息处理等功能有机结合起来实现用户要求的过程。

3) 资金流:是指在营销渠道(或者供应链)成员间随着商品实物及其所有权的转移而发生的资金往来流程。

4) 信息流:是指交易过程涉及的各个主体(包括营销渠道成员、供应链成员、商品)以及各种商务活动要素(商流、物流、资金流)相关的信息汇聚,以确保整个商务活动的有效进行。

商流、物流、资金流和信息流的关系可以理解为:商流是动机和目的、资金流是条件、信息流是手段、物流是终结和归宿。物流是电子商务的基础,虽然可以用各种信息手段提升物流的效率、降低物流的成本、透明化物流的过程,但毋庸置疑,相比较于几乎可以完全互联网化的商流、资金流和信息流而言,物流是“重”的。或许正是因为这个原因,国家从产业引导的角度也进行了相关规划,比如科技部发布的国家科技支撑计划项目指南(2015)中的信息产业与现代服务产业专题中,专门针对(跨境)电商中的物流及物流终端综合服务给出专项指南;而国务院公开发布的《国务院关于积极推进“互联网+”行动的指导意见》(国发〔2015〕40号)也专门将高效物流纳入“互联网+”的重点行动规划中。

电子商务涉及的主体有商家(Business)、普通消费者(Customer)、家庭消费者(Family)、产品(Production)、市场(Marketing)、政府(Government)、制造厂商(Manufacturers)等,根据商务交易参与角色的不同可以将电子商务分为若干类,以下罗列了一些主流的电商模式:

1) B2B:泛指企业与企业之间通过互联网进行产品、服务及信息的交换。如阿里巴巴、中国制造网、敦煌网、慧聪网。

2) B2C:泛指企业与个人之间通过互联网进行产品、服务及信息的交换。这类电商模式有两种具体的做法,一种是在电商平台如淘宝、京东、苏宁上开设线上商城,利用电商平台的渠道和品牌进行商务活动;另外一种开设专门的垂直电商平台(往往自营)进行商务活动。

3) B2G: B2G电子商务模式即“商家到政府”,是企业与政府之间通过网络所进行的交易活动的运作模式,如电子通关、电子报税。

4) B2M: 泛指企业对市场的商务模式, 是电子商务公司以客户需求为核心而建立起的营销型站点, 并通过线上和线下多种渠道对站点进行广泛推广和规范化导购管理, 从而使得站点成为企业的重要营销渠道。

5) B2F: 泛指企业对家庭的商务模式, 是商务机构按交易对象分类, 把每个人分类于家庭这个单位之中, 并以便捷的购物方式来引导消费; 通过一站式服务和高效免费的配送、安全可靠的现金交易来赢取市场位置。这种形式的营销模式一般以目录 + 网络销售为主。

6) C2B: 泛指消费者到企业的商务模式, 这一模式改变了原有生产者(企业和机构)和消费者的关系, 是一种消费者贡献价值, 企业和机构消费价值的模式。通常情况为消费者根据自身需求定制产品价格, 或主动参与产品设计、生产和定价, 产品、价格等彰显消费者的个性化需求, 生产企业进行定制化生产。

7) C2C: 就是个人与个人之间的电子商务。比如一个消费者有一台电脑, 通过网络进行交易, 把它出售给另外一个消费者, 此种交易类型就称为 C2C 电子商务。

8) C2G: 消费者对行政机构间的电子商务, 指的是政府对个人的电子商务活动。这类的电子商务活动目前还没有真正形成。然而, 在个别发达国家, 如在澳大利亚, 政府的税务机构已经通过指定私营税务, 或财务会计事务所用电子方式来为个人报税。这类活动虽然还没有达到真正的报税电子化, 但是, 它已经具备了消费者对行政机构电子商务的雏形。

9) G2C: 是指政府与公众(Citizen)之间的电子政务, 是政府通过电子网络系统为公民提供各种服务。电子政务的目的不仅是政府给公众提供方便、快捷、高质量的服务, 更重要的是可以开辟公众参政、议政的渠道, 畅通公众的利益表达机制, 建立政府与公众的良性互动平台。

10) G2G: 是一种政府对政府的电子政务应用模式, 是电子政务的基础性应用。

11) M2C: 是指制造厂商直接对消费者提供自己生产的产品或服务的一种商业模式, 特点是流通环节减少至一对一, 销售成本降低, 从而保障了产品品质和售后服务质量。

12) P2C: 泛指产品从生产企业直接送到消费者手中的商务模式, 中间没有任何的交易环节, P2C 把老百姓日常生活当中的一切密切相关的服务信息, 如房产、餐饮、交友、家政服务、票务、健康、医疗、保健聚合在平台上, 实现服务业的电子商务化。

13) O2O: O2O 是 Online To Offline 的缩写, 指的是将线下商务的机会与互联网结合在一起, 让互联网成为线下交易的前台, 这是有别于上述分类标准的一种电子商务模式(其实是一种商业模式, 后文会有更为详细的介绍)。

艾瑞咨询统计数据显示, 2014 年中国电子商务市场交易规模 12.3 万亿元, 同比增长 21.3%。其中 B2B 仍是电子商务的主体, 占比 73.4%(中小企业 B2B 电子商务占 50%, 规模以上企业 B2B 电子商务占 23.4%), 网络购物同比增长 48.7%, 占 22.9%(其中 B2C 占 45.85%, 同比增加 68.7%; C2C 占 54.2%, 同比增加 35.2%, B2C 市场将继续成为网络购物行业的主要推动力), 在线旅游占 2.3%, 同比增长 27.1%, 本地生活服务 O2O(餐饮、婚庆、休闲娱乐、亲子、美容、美护等细分行业)占 1.4%, 同比增长 42.8%, 共同促进电子商

务市场整体的快速增长。(摘自 <http://www.iresearch.com.cn/>)

事实上,本文无法罗列所有的电商模式,因为创新的电商模式总会被不断地创造出来。以 B2B2C (Business to Business to Customer) 为例,它是近年流行起来的一种商务模式,该模式将 B2B 和 B2C 有效结合起来,第一个 B 指的是商品或服务的供应商,第二个 B 指的是从事电子商务的企业,C 则是表示消费者,在第二个 B 构建的统一电子商务平台购物的消费者。

某种意义上而言,传统的电子商务仅关注商品交易相关的内容,比如商家考虑的是:卖什么?货在哪里?怎么卖?客户在哪里?如何引流?换句话说,传统电商是可以不关心产品是由谁以及在哪里生产这件事情的,而最近衍生出的若干电子商务新模式,比如前面提到的 C2B,则分明表达了一个声音:由用户端驱动进行产品的设计和生产。这个思路的革命性在于:电子商务的关注点应该从传统的营销向产品的设计、制造和生产延伸,如是就可以将“产品 \leftrightarrow 用户”形成一个完备的闭环,而所有环节(设计、生产、制造、营销、仓储等)都可以在这个闭环里实现水平价值和垂直价值的有效整合,并联动所有相关产业链的优化和调整。换句话说,电子商务的迅猛发展能够带动和促进传统制造业的发展和进步。

13.3.2 移动电子商务

移动通信技术的发展引发了移动互联网的发展,智能终端的普及以及资费的下降进一步促进和推动了移动互联网的发展。移动互联网的3个重要特征是3A 便捷性(AnyTime、AnyPlace、Anything)、智能感知性(智能终端本身可以更广义地采集反映终端的环境)和个性化(终端、网络、内容及应用的个性化),无线宽带技术的不断发展进一步强化了这三方面的特征。移动互联网向各个领域的普及和渗透,在改变和改善人们生活工作的同时,也逐步成为人们生活、工作的一部分。

在移动互联网市场不断扩大的背景下,如何充分利用移动互联网的这一特征进行应用创新、模式创新、技术创新是工业界、学术界普遍关心的重要课题。移动电子商务就是利用手机、PDA 等无线终端进行的电子商务活动。

移动终端设备以其便利性、功能齐全等特性成为未来人们商务和生活的重要应用工具,移动电子商务也成为继传统互联网电子商务后又一令人关注的领域。艾瑞咨询最新统计数据 displays,2014 年中国移动购物市场交易规模为 9297.1 亿元,年增长率达 239.3%,远高于中国网络购物整体增速(2014 年中国网络购物市场交易规模为 28145.1 亿元,较去年同期增长 49.8%)。艾瑞预测未来几年中国移动购物市场仍将保持较快增长,2018 年移动购物市场交易规模将超过 4 万亿元。在 PC 端与移动端份额占比方面,2014 年中国移动购物交易额在中国网络购物整体市场中占比 33.0%,较 2013 年增长近 19 个百分点。艾瑞预计移动端交易占比在未来几年继续上升,2016 年将超过 PC 端网购交易占比,成为中国网民网购的重要选择。

(摘自 <http://www.iresearch.com.cn/>)

移动电子商务产业链中涉及的5个主要对象包括用户、商户、网络提供商、应用提供商和设备提供商。按照移动电商产业链中主导者的不同,移动电商的商业模式可分为如下几种:

1) 电信运营商主导的移动电商:电信运营商是网络平台的提供者和支撑者,主导的是一种“通道+平台”的商业模式。

2) 传统电子商务提供商主导的移动电商:主导的是一种“品牌+运营”的商业模式,如淘宝、当当、Amazon、eBay。

3) 设备提供商主导的移动电商:设备提供商是终端设备提供者,主导的是“设备+服务”的商业模式,如苹果公司的App Store。

4) 新兴移动电子商务提供商主导的移动电商:新技术结合各式各样的“创新应用”,通过应用来吸引用户,引导用户的消费模式。在应用为本的时代,以应用创新为导向无疑会吸引用户眼球,增加用户的黏性,这是此模式的优势所在。

上述4类移动电子商务主导方是未来移动电子商务行业发展的关键力量,其业务发展的策略代表了未来移动电商发展的基本方向。

目前,我国移动电子商务随着移动通信技术的进步取得了长足的发展,也引起了我国电子商务界的广泛关注。但是,和移动电子商务发达国家相比,我国移动电子商务还处于发展的初级阶段,还存在一些亟待解决的问题。就当前的情况看来,我国移动电子商务的发展存在的主要问题有:完善移动电子商务的安全保障问题、完善相关法律法规、建立健全信用机制、加强身份认证、减少移动终端丢失、完善移动电子商务商业模式、完善移动支付体系等。

总体而言,电子商务在压缩中间环节、化解产能过剩、促进企业发展、增加就业岗位方面都大有裨益,同时作为一个撬动整个产业链的杠杆,也能够通过类似于C2B这样的电商模式,让最终消费者参与到产品的设计、研发和制造中,这不仅使企业能够生产出更满足用户个性化需求的产品,更重要的意义在于,作为产品的制造厂商可以实时、快捷地了解消费舆情、个性偏好、市场趋势等,从而提高制造厂商的竞争力和生命力提供有价值的基础支撑。经济学家吴晓波如是说:中国经济赶超美国,必须用信息化再造中国制造模式,“电子商务+专业公司+小制造”模式将大行其道,也是这个道理。

13.3.3 跨境电子商务

跨境电子商务是指分属不同关境的交易主体,通过电子商务平台达成交易、进行支付结算,并通过跨境物流送达商品、完成交易的一种国际商业活动。作为推动经济一体化、贸易全球化的技术基础,跨境电子商务具有非常重要的战略意义。对于发展中的中国而言,跨境电商除了能够丰富经贸活动、拓宽企业销售渠道、实现企业互赢、便捷人们的消费外,还能够打造中国自主品牌、提高中国品牌海外影响力、加强外汇储备、提升中国制造在全球的地位等方面起到极大的推动作用。

2014年3月5日,李克强总理在政府工作报告中指出:“……要稳定和完善出口政策,加快通关便利化改革,扩大跨境电子商务试点……引导加工贸易转型升级,支持企业打造自主品牌和国际营销网络,发展服务贸易和服务外包,提升中国制造在国际分工中的地位……”2015年的两会工作报告中,总理再次提及跨境电子商务的推动和发展战略。

跨境电子商务的本质仍然是通过物流、资金流和信息流建立起卖家到买家、工厂到终端用户的交易。整个电子商务涉及的环节包括(但不限于)市场咨询、广告投放、询价议价、商品搜索、产品舆情、商检、通关、售后维保、一般仓储、质检、税务、支付、法律法规、合同管理、信用担保、纠纷仲裁、一般物流、报关通关、跨境物流等。围绕不同环节的细分需求,产生若干专门面向跨境电子商务的服务商,比如一般贸易服务商(为一般贸易提供普适贸易服务)、第三方电商服务商(为一般贸易活动提供增值业务)、第四方电商服务商(为细分需求进行二次开发的软件商)。所有这些服务商提供的跨境电商服务模式有(但不限于)四类,参见表13-2。

表 13-2 跨境电商主要服务模式

序号	服务模式	服务内容	典型企业
1	自建网站自主销售	直接销售给海外当地客户或经销商	兰亭集势, DX.com, 环球易购
2	建立独立网站的第三方销售平台	小商品的展示、交易、结算服务	阿里速卖通, 敦煌网
3	传统的外贸服务	提供通关、商检等基础服务	一达通, 贸易通
4	其他	物流仓储、支付、软件等	Paypal, 4PX, 出口易

我国跨境电子商务主要分为 B2B、B2C、C2C 三种商务模式。B2B 模式下,企业运用电子商务以广告和信息发布为主,成交和通关流程基本在线下完成,本质上仍属于传统贸易,已纳入海关一般贸易统计。B2C 及 C2C 模式下,我国企业(或个人)直接面对国外消费者,以销售个人消费品为主,物流方面主要采用航空小包、邮寄、快递等方式,其报关主体是邮政或快递公司,目前大多未纳入海关登记,后者一般也称为小额跨境电商。相比较而言:

1) B2B 模式的物流成本远小于 B2C 模式的物流成本,且 B2B 模式的跨境物流已经成熟,跨境通关效率容易提高。

2) B2C 模式尤其是 C2C 模式的跨境物流除了具有典型的总体批次大、单次盈利小,累计物流成本大的特点外,通常其商检效率低,或者直接回避海关登记,成为跨境电商中的“灰色”地带。

3) 基于 B2B 和 B2C 基础上发展起来的 B2B2C(前文提及),即“企业→小企业→客户”,这种模式兼具 B2B 和 B2C 的特点,是目前中国跨境电商中的新生趋势。

4) 经过多年的发展,中国的“一般贸易”(生产+跟单+仓储+质/商检等)模式已经渐趋成熟,但是在贸易服务能力方面仍有欠缺。

跨境电商涉及的产业链涉及角色、环节太多,几个共性的问题异常突出,分别是基础信息保障、基础业务协同、业务系统集成。

(1) 基础信息保障问题

跨境电商涉及的相关环节以及因此而产生的各类数据组成了跨境电商活动中的基础信息,比如采购商信息(信用、资金能力等)、工厂信息(如信誉、产能、资金能力)、订单信息(历史交易信息、采购数量、种类等)。这些基础信息的可信是保障电商活动健康运作的重要基础,特别值得一提的是,在小额跨境电商场景下(小额贸易相对于大额贸易而言),上述信息的获得和可信度评估更加困难。

另一方面,(跨境)电子商务所涉及的物流、资金流和信息流本质上都是面向业务协同和集成的数据流,如何使“源头数据获取→目标应用涉及的业务流程使用→存档”整个流程中的数据可信也是保障电商活动健康运作的重要基础。这意味着在网络安全、信息安全保障体系的基础上,对于基础信息数据在整个数据生命周期内的质量监管非常重要,比如:数据的采集是否有误,数据的使用是否不当引发篡改。

(2) 基础业务协同问题

信息化推动和深化了常规贸易的快速、高效进行,(跨境)电子商务本质上是一种四流合一的贸易。资金流和信息流本身都是一种互联网意义上的“线上行为”,互联网的“互联精神”已经为线上行为提供了一种可行、可信的通信基础。而物流是一种典型的“线下行为”,这种“线下行为”直接对应于质检、税务、物流等相关单位在商务层面的对接,在跨境电商意义下的基础业务除此之外还应当包括因为“跨境”而增加的与商检、海关等单位在商务层面的对接。

数据规范化是确保在商务层面与各相关单位进行高效对接,但是涉及的基础业务单位在业务逻辑、政策理解、区域分布等众多方面的差异性和多义性给数据的规范化带来了很大的挑战。比如,对于通关产品而言,尽管已有既定的分类准则,但是产品描述本身的多维性以及不同维度下的产品分类在线下的税务、商检上均有很大的不同,直接影响通关的效率。

(3) 业务系统集成问题

电子商务的本质是通过信息化的注入,通过信息流、物流、资金流的高效处理以快速嫁接生产原厂与终端客户,从而达到两端的快速交互,比如更快的物流将产品更快地送到终端客户、更快的信息反馈将终端客户的消费行为、偏好等反馈至原厂。显然其间涉及的业务系统和行业标准很多,比如在线支付、安全保障体系、通信协议、支付标准、银行系统,因此提高与网络交易相关业务的集成度是提高电商效率的重要保障。

数据集成是业务集成的基础,而在网络交易中涉及的数据具有极强的分布性、异构性及多模式性等特点,比如在线支付的风险评估涉及的数据至少包括:支付方的网络环境、操作系统环境、区域信息、通信协议等。这意味着业务集成的目标达成必须有效融合、集成异构数据,并在此基础上,提供有效的数据评估手段和策略。

13.3.4 应用提示

在“互联网+”作为国家战略、行业趋势和潮流的同时,把(跨境)电子商务以三驾马

车的角色纳入国家战略,可能的原因或许在于(不限于):

1) 作为一种应用模式,电子商务在丰富和繁荣了消费市场、顺应了人的消费偏好、方便了人的行为习惯的同时也改造了传统的生活方式,并且向生活质量更高的方向前行。

2) 作为一种营销渠道,电子商务在拉近和强化了买卖关系、诱发了人的消费欲望、拓展了商家盈利空间的同时也改进了传统的商业模式,并且向多边互利共赢的方向推进。

3) 作为一种供需杠杆,电子商务在联动和撬动了制造生产、激发了企业创新能力、倒逼了产品质量改进的同时也推进了产业结构的优化,并且向综合国力提高的方向前行。

当然,上述目标和愿景的达成需要多边力量的共同努力,就计算机科学与技术而言,针对电子商务涉及的各个环节提供技术服务、进行技术攻关从而助力细分目标的达成,应当是每一位计算机工作者的使命,也是机遇。作为一种应用示例,本文尝试罗列(但不局限于)一些大数据(技术)在电子商务领域的若干细分应用场景。

(1) 物流

物流(配送)是电子商务中的重要一环,也是必须线下进行的(除非是商品的属性使之可以在网上传送,比如虚拟产品;或者是根本无须进行物流的产品,比如房地产),因此物流效率的提高对于商流目标的达成具有重要的意义。

如前所述,物流的核心是仓储。传统的仓储被视为无附加价值的成本中心,而现代仓储作为连接供方和需方的桥梁,被视为企业成功经营的关键。从供方来看,作为流通中心的仓储关注的是高效的流通加工、库存管理、运输和配送等活动;从需方来看,作为流通中心的仓储关注的是以最大的灵活性和及时性满足各类顾客的需要。因此,精准的仓储管理能够有效控制和降低流通和库存成本,是助力企业保持优势的关键。显然,此处的仓储管理不仅仅是传统意义上对仓库物品的保管管理、作业管理、安全管理、流程管理等,更重要的在于和其他环节的有效、实时联动,比如根据市场需求制定弹性的生产计划、根据市场的实时销售动态调度和配送,前者涉及弹性制造的话题(13.4节专门介绍),后者涉及对市场销售数据的实时采集、存储、分析建模和快速策略响应,是典型的大数据应用场景。

将自己定位成“大自然搬运工”的农夫山泉,在全国有十多个水源地。农夫山泉把水灌装、配送、上架,一瓶超市售价2元的550ml饮用水,其中3毛钱花在了运输上,如何根据不同的变量因素来控制自己的物流成本,成为问题的核心(另外一个问题是,如何平衡生产和销售,此处不赘述)。基于上述场景,SAP团队和农夫山泉团队提供了一个大数据解决方案的思路:在数据收集层次,除了采集企业内部数据外,还将很多外部数据纳入进来,比如高速公路的收费、道路等级、天气、配送中心辐射半径、季节性变化、不同市场的售价、不同渠道的费用、各地的人力成本甚至突发性的需求(比如某城市召开一次大型运动会)……通过对农夫山泉内部数据和外部数据的有效采集和整合,并进行数据建模和高性能计算,为物流配送、企业生产提供辅助决策意见,这是一个典型的大数据应用案例,该解决方案的特色在更多数据的采集以及更高计算性能的保障(这个是通过SAP HANA数据库的实时计算加以保障的)。

(2) 广告营销

所谓广告,指的是为了某种特定的需要,通过一定形式的媒体,公开而广泛地向公众传递信息的宣传手段,是商品生产者、经营者和消费者之间沟通信息的重要手段。而互联网营销指的是在互联网平台上实现辅助营销目标的市场营销方式。随着互联网的不断发展和深入,互联网营销的策略和手段也多有不同并在不断发展中,比如 SEO、社会网络营销、精准营销,每一类名词背后都代表了一种营销理念。

SEO 指的是搜索引擎优化 (Search Engine Optimization),是指基于对主流搜索引擎 (比如谷歌、百度) 数据获取和搜索结果排名方法的了解,对企业自营网站进行内部及外部的调整优化,使得以某些关键字 (词) 在搜索引擎的搜索排名更加靠前,从而为企业自营网站获得更多流量,进而达成网站销售及品牌建设的目标。SEO 的营销理念主要在于:用户购买某些物品时一定会在搜索引擎中检索,并有很大可能会点击进入排名靠前的网站。社会网络营销指的是将虚拟社交平台 (比如微博、微信) 作为媒介进行广告的推送、公关、推广以及与潜在客户的互动,是一种典型的整合营销行为,偏重于口碑效应的传播。精准营销指的是在 (对目标客户) 精准定位的基础上,将合适的产品、通过合适的渠道、在合适的时机推送给合适的人,并建立双向的透明沟通。

精准营销理念的奠定诱发了商家对潜在用户立体化刻画的需求:客户在哪里?有何偏好?自然属性如何?价值度和忠诚度如何?通过怎样的渠道进行?所有这些问题事实上是精准营销面临的最大“痛点”,而这也是大数据可以发挥用武之地的重要战场。比如在搜索引擎里搜索的关键字可以反映这个账号主体 (或者 IP 地址) 的需求;社会网络中发布的言论可以反映账号主体的价值观偏好 (甚至行动轨迹、短期需求等);在电商网站上的购物行为可以反映这个账号主体的消费偏好、能力和消费习惯;电信运营商的数据库中则富含着每个账号主体的活动运行规律、消费能力、通信习惯、朋友圈;导航地图软件则记录着主体的出行规律;信用卡消费记录不仅记录下每个人的消费偏好,还能反映信用卡主人的区域活动范围;公交卡用于支付交通费用的同时,事实上也记录了公交卡的出行范围等。对上述数据的有效整合以及对人的立体化描述,对于商家的精准营销而言,无疑大有裨益。当然,如何将上述数据集成一个数据平台上,需要合适的应用模式和商业模式支撑 (上述的这些数据散布在各个利益主体的数据库里)。事实上,在实际应用中,某一个数据源的数据就可以在某个角度对人进行刻画,就已经具有了价值。另外一个需要考虑的问题是如何将虚拟网络平台的虚拟身份和反映物理社会的真实身份关联起来 (对应一个物理的“人”),这需要技术手段加以解决,此处不加议论。

上面的简单分析至少表明了一个事实:在面向精准营销的人的立体画像的应用目标下,开展系统设计与实现时,除了需要注重数据的混杂性、注重数据的交叉复用、注重数据的相关性分析、注重数据的广度外,还要注重数据 (源) 的质量等,所有这些均是大数据思维。

13.4 工业互联网

13.4.1 基本概念

工业互联网是2015年3月5日李克强总理在政府工作报告中提出“互联网+X”战略的同时,提及优先发展和支持的战略规划。工业互联网是中国版的“工业4.0”,也是“互联网+”的重要战场。为了实施制造强国战略,2015年5月8日,经李克强总理签批《中国制造2025》,部署全面推进实施制造强国战略。《中国制造2025》提出,坚持“创新驱动、质量为先、绿色发展、结构优化、人才为本”的基本方针,坚持“市场主导、政府引导,立足当前、着眼长远,整体推进、重点突破,自主发展、开放合作”的基本原则,通过“三步走”实现制造强国的战略目标:第一步,到2025年迈入制造强国行列;第二步,到2035年我国制造业整体达到世界制造强国阵营中等水平;第三步,到2049年,我国制造业大国地位更加巩固,综合实力进入世界制造强国前列。

没有用互联网工业这样“互联网+”的说法,或许是因为这个领域(行业)太重,互联网化的难度太大,而采用了“X+互联网”的说法。在工业互联网这个概念因为“互联网+”而广为流行之前,与此目标相同的战略是两化融合:江泽民总书记在十六大报告中提出“……以信息化带动工业化、以工业化促进信息化……”,胡锦涛总书记在十七大报告中提出“……促进信息化与工业化融合,走新型工业化道路……”,李克强总理在2014年政府工作报告中提出“……促进信息化与工业化深度融合……”。

国务院在2015年7月1日公开发布的《国务院关于积极推进“互联网+”行动的指导意见》(国发〔2015〕40号)再次提出将协同制造纳入“互联网+”的重点行动规划中:“推动互联网与制造业融合,提升制造业数字化、网络化、智能化水平,加强产业链协作,发展基于互联网的协同制造模式。在重点领域推进智能制造、大规模个性化定制、网络化协同制造和服务性制造,打造一批网络化协同制造公共服务平台,加快形成制造业网络化产业生态体系”。

13.4.2 笑脸曲线

制造业是国民经济的主体,是立国之本、兴国之器、强国之基。打造具有国际竞争力的制造业,是我国提升综合国力、保障国家安全、建设世界强国的必由之路。

宏碁集团创办人施振荣先生提出并得到制造业同行认可的笑脸曲线描述了生产制造涉及的各个环节的附加值收益分布,参见图13-2。

从图13-2可以看到,在整个产业(价值)链中,制

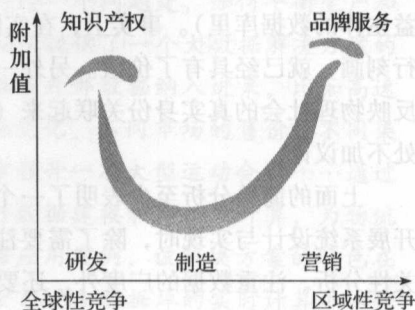


图 13-2 笑脸曲线

造环节的附加值最低,附加值最丰厚的区域集中在价值链两端的研发和营销(研发的附加值获益体现在知识产权,属于全球性竞争;营销的附加值获益体现在品牌服务,属于区域性竞争)。企业如果要获得更多的附加值,就必须向两端延伸,要么向上游的零件、材料、设备及科研延伸,要么向下游营销端的销售、传播、网络及品牌延伸。总体而言,愈向两边走,企业获得的附加值就越多。

笑脸曲线有两个要点:一是可以找出附加值在哪里;二是关于竞争的形态。华为的人力资源配置一直呈“研发和市场两边高”的“笑脸曲线”,这或许是华为一直立于不败之地的一个原因。20世纪90年代,国际品牌虽然有技术优势,但其价格远高于华为,而且其服务速度也很难跟上。华为以其特有的“狼性文化”在市场上攻城略地,大规模的营销人员确定了华为在市场上的优势,他们为客户提供快速而周全的贴身服务,这是其向营销端延伸并通过服务扩张更多附加值获益的策略。经过多年的积淀,华为作为一个国际品牌,已经举世瞩目,这是通过品牌站到了附加值获益的顶端。向上游延伸方面,研发和创新是华为的核心竞争力,并已经成为华为基因的一部分。截至2014年年底,华为累计获得专利达到38825件,其中90%以上为发明专利。根据联合国机构世界知识产权组织公布的报告,华为2014年申请国际专利3442件,在全球企业中排名第一。

我国制造业经过几十年的发展,已经建成了门类齐全、独立完整的产业体系,有力推动了工业化和现代化进程。然而,与世界先进水平相比,我国制造业仍然大而不强,大多附加值的获益集中在笑脸曲线的底层,在自主创新能力、资源利用效率、产业结构水平、信息化程度、质量效益等方面差距明显,转型升级和跨越发展的任务紧迫而艰巨。而全球制造业格局的重大调整以及国内经济发展环境的重大变化作为主要的外因和内因也使得中国制造产业结构调整、转型和升级更加迫切。

我国制造业面临发达国家和其他发展中国家“双向挤压”的严峻挑战:

(1) 美国制造业回归

长期以来,作为老牌科技强国的美国一直处于笑脸曲线的附加值最高端,而将附加值低端的制造以外包的形式放在发展中国家进行。但是国际金融危机发生后,美国与其他发达国家一道纷纷实施“再工业化”战略。

在美国,共和、民主两党都不约而同地把刺激制造业作为本党的竞争纲领和政治战略(这或许也是两党大选竞争中难得的相同点)。2009年奥巴马政府签署《美国振兴和再投资法案》刺激制造业;2010年签署《美国制造业促进法案》通过暂停或减少进口关税降低成本;2011年宣布《先进制造合作计划》把行业、学校和联邦政府联合在一起,并在科研上投入5亿美元,希望带来美国制造业的复兴;2012年奥巴马政府再次发表国情咨文,提出从税收政策上刺激制造业的提议计划;2012年美国大选奥巴马再次打振兴制造业的牌;2013年首次国情咨文讲话中又提出了振兴美国制造业的计划……

美国如此高调提出振兴制造业、制造业回归的原因或许有很多,其中一个应该是通过刺

激制造业,提高本国的就业率。比如,美国制造业联盟2011年发布的报告表明86%的选民支持国家刺激制造业发展;波士顿咨询公司在2013年11月发布的咨询报告表明80%的人愿意出更多钱买美国制造的商品,其中93%的人的理由是要把工作留在美国。

另外一个原因或许是美国的科技发展优势使然。美国每年的开发研究投资大概有2/3都和制造业有关,新技术的运用一方面形成了新的行业,同时新技术提供高效的管理,提高了整个制造业的设计速度。同样是因为科技的发展,美国的能源成本逐步降低,比如美国页岩气的快速发展对美国乃至全球的能源格局形成了不容忽视的影响(此处不赘述),直接降低了制造企业的制造成本。加之其他发展中国家用工成本的逐步提高及物流费用的不断上扬,使得原先将低端制造外包出去的重要原因瓦解。

总之,美国制造业回归的事实是:美国制造业总体从2011年的1.73万亿美元增长到2012年的1.87万亿美元,增幅8.09%(同期的美国总体经济增长为2.5%),美国生产的产品占全球产品的21%,2014年4~7月,制造业的就业人数稳定在接近1200万的水平(美国统计局的数据),据波士顿咨询公司预测,到2020年,制造业和出口的发展将给美国增加250万~500万个就业机会。

(2) 欧盟新工业革命及德国工业4.0国际战略的出台

欧债危机充分暴露了欧盟“去工业化”致使部分成员国抗危机能力不足的弱点。欧盟经济持续低迷的现实促使欧盟重新检讨和调整自己的产业政策,新工业革命已经被提到了欧盟产业政策调整的议事日程,欧盟提出的新工业革命就是要让工业重新回到欧洲,把发展工业实体经济作为重点。

2010年,欧盟委员会制定了《全球化时代的统一产业政策》文件,提出了未来工业政策的基本框架,内容涉及改善产业环境、强化内部市场、工业创新政策、国际资本化和促进工业现代化等几大方面。2012年10月,欧盟委员会发表了《强大的欧盟工业有利于经济增长和复苏》的工业政策通报,确立了欧盟工业的核心地位,提出通过新工业革命扭转欧盟工业比重下降趋势,到2020年将欧盟工业增加值在国内生产总值中所占比例由目前的16%提升至20%。欧盟所采取的这一系列重大举措不仅表明了欧盟产业政策调整的战略方向正在发生重大变化,还体现了欧盟寻求通过推行新工业革命来促使经济复苏和创造更多就业机会的目的。

当然,欧盟的再工业化不是简单的重复,而是由先进创新技术带动的新型工业革命。强调技术创新、结构改革,改变碳氢化合物为主的能源结构,大力推进新的生产方式,如机器人、数字技术、先进材料、可循环能源。

而老牌制造强国德国在《高技术战略2020》确定的十大未来项目中,将“工业4.0”纳入国家战略,以提高德国工业的竞争力,在新一轮工业革命中占领先机。工业4.0是以智能制造为主导的第四次工业革命,通过革命性的生产方法,利用信息物理系统(Cyber Physical System, CPS)将生产中的供应、制造及销售信息数据化、智慧化,最后达到快速、有效及个性化的产品供应。分为两大主题:一是“智能工厂”(智能化生产系统及过程,以及网络化分

布式生产设施的实现);二是“智能生产”(生产物流管理、人机互动以及3D技术在工业生产过程中的应用等)。

鉴于德国制造强国的示范性,工业4.0已经成为各国政府角力制造产业的样本,包括我国政府提出的工业互联网事实上就是中国版的工业4.0。关于“工业4.0”的话题在13.4.3节专门介绍,此处不赘述。

(3) 中国人口红利的丧失及国内发展环境的重大变化

一方面,随着新型工业化、信息化、城镇化、农业现代化的同步推进,超大规模内需潜力不断释放,为我国制造业发展提供了广阔空间。各行业新的装备需求、人民群众新的消费需求、社会管理和公共服务新的民生需求、国防建设新的安全需求,都要求制造业在重大技术装备创新、消费品质量和安全、公共服务设施设备供给和国防装备保障等方面迅速提升水平和能力。

另一方面,我国仍处于工业化进程中,与先进国家相比还有较大差距,主要依靠资源要素投入、规模扩张的粗放式发展方式难以为继,总体水平低、大而不强、创新能力弱、产能结构性过剩。而且我国人工红利的丧失,使得中国的劳动力等生产要素成本不断上升的同时,印度、越南、墨西哥等发展中国家以更低成本优势成为接纳发达工业国家产业转移的新阵地,加之前面提及的发达国家的重振制造业给我国制造业在承接产业转移、技术进步与出口等方面带来更为严峻的挑战。

总体而言,推进制造强国建设,必须着力解决下述问题(不限于):一是我国制造业自主创新能力弱,关键核心技术与高端装备对外依存度高。二是以企业为主体的制造业创新体系不完善,产品档次不高,缺乏世界知名品牌。三是资源能源利用效率低,环境污染问题较为突出。四是产业结构不合理,高端装备制造业和生产性服务业发展滞后。五是信息化水平不高,与工业化融合深度不够。六是产业国际化程度不高,企业全球化经营能力不足。

事实上,我国从十六大报告中提出“以信息化带动工业化、以工业化促进信息化”开始,一直强调两化融合战略和两化深度融合战略,2015年的政府工作报告又再次把此战略升格为“互联网+”时代的工业互联网战略,本质的目标都是一个:产业结构调整、转型和升级,以实现中国的制造强国梦。

13.4.3 工业4.0

“工业4.0”这个概念是在2013年4月的汉诺威工业博览会上正式推出,后来上升为德国政府的国家战略,其目的是提高德国工业的竞争力,在新一轮工业革命中占领先机。该战略已经得到德国(事实上也包括德国以外的其他国家)科研机构 and 产业界的广泛认同。

之所以用“工业4.0”这样的称呼,是因为在历史上已经发生了三次工业革命:

(1) 第一次工业革命

一般认为的第一次工业革命是从18世纪50年代到19世纪40年代前后,这个时间段跨越了我国大清王朝的乾隆(1735—1795)、嘉庆(1796—1820)、道光(1821—1850)共三个帝

王时代。

一般用“蒸汽（机）时代”作为第一次工业革命的标签，其主要的原因在于英国发明家詹姆斯·瓦特（James Watt）在1776年改良的蒸汽机（经过一系列改良后，使之成为万能的原动机）将人类从手工作业时代直接带入机器作业时代。这不仅是一次技术变革，更是一场深刻的社会变革，比如工厂制代替了手工工场、机器代替了手工劳动，而从社会关系来说，第一次工业革命使依附于落后生产方式的自耕农阶级消失了，工业资产阶级和工业无产阶级形成和壮大起来。同时，第一次工业革命大大加强了世界各地之间的密切联系，改变了世界的面貌，最终确立了资产阶级对世界的统治地位，率先完成了工业革命的英国，很快成为世界霸主。

事实上，在瓦特发明蒸汽机以前，就有发明家改进当时手工工场的作业方式，以提高生产效率。在棉纺织行业，人们先是发明了一种叫飞梭的织布工具，大大加快了织布的速度，也刺激了对棉纱的需求。18世纪60年代，织布工詹姆斯·哈格里夫斯发明了珍妮手摇纺纱机（珍妮是其女儿的名字）。珍妮手摇纺纱机事实上是对旧式纺车的一个技术改进：用一个纺轮带动多个（最开始的原始是八个）纱锭，借此提高生产率。虽然还是使用人力进行，但是珍妮手摇纺纱机的出现，使大规模的织布厂得以建立。因此，“珍妮手摇纺纱机”的发明被认为是第一次工业革命的开端。

1769年，英国人理查德·阿克莱特发明了水力纺纱机，用水力代替人力作动力进行生产（纺出的纱坚韧结实，但比较粗，且一旦天气干旱，河流枯竭，就不能运转），促进了英国工业革命的开展。第一次工业革命中颠覆性的事件发生在1785年，瓦特改良的蒸汽机用于改良纺纱机，直接推动机械化大生产时代的到来，同时蒸汽机还被用于交通运输等场景……

第一次工业革命的技术发明主要是工匠们的经验结果，科学起了辅助作用（技术上的发明由于受到科学的指引而更加迅速发展），而由工作机、传动机、动力机所组成的机器系统的出现使得机械化大生产成为可能，同时机器使得劳动进一步分工，产生了工匠和工程师的分工，为自然科学的发展和应用提供了新的基础和条件，而技术发明家已成为一种独立的社会职业。

（2）第二次工业革命

一般认为的第二次工业革命是从19世纪60年代至19世纪末20世纪初前后，这个时间段跨越了我国大清王朝的咸丰（1851—1861）、同治（1862—1874）、光绪（1875—1908）、宣统（1909—1911）共四个帝王时代以及后来的民国（1912—1949）。

一般用“电气时代”作为第二次工业革命的标签，其主要的原因在于使用石油和汽油的内燃机的发明和推广应用，使得动力工业被彻底改革，将人类从蒸汽机时代直接带入电气时代。与第一次工业革命相比，第二次工业革命具有几个典型的特点：

1) 自然科学同工业生产紧密结合起来，科学在推动生产力发展方面发挥更为重要的作用，而不像第一次工业革命大多源于工匠的实践经验。

2) 第二次工业革命是在几个先进的资本主义国家同时进行的, 其规模更加广泛, 发展也比较迅速, 而不像第一次工业革命仅在英国进行。

3) 有些主要资本主义国家的两次工业革命是交叉进行的 (主要的原因是诸如日本这样的资本主义国家尚未完成第一次工业革命, 对它们来说, 两次工业革命是交叉进行的)。

第二次工业革命极大地推动了社会生产力的发展, 对人类社会的经济、政治、文化、军事、科技和生产力产生了深远的影响。内燃机的应用解决了交通工具的发动机问题, 推动了交通工具的发展, 出现了早期汽车和飞机, 这也直接推动了石油开采业和石油化工工业的发展, 1867 年诺贝尔发明了炸药, 80 年代发明了无烟炸弹, 同时能够从石油中提炼出氨和苯……而科学技术的进步也带动了电讯事业的发展, 19 世纪 70 年代, 美国人贝尔发明了电话, 90 年代意大利人马可尼试验无线电报取得了成功, 这都为迅速传递信息提供了方便。世界各国的经济、政治和文化联系进一步紧密。

(3) 第三次工业革命

一般认为的第三次工业革命从二战至今, 我国从战火年代走进了和平年代, 也因为中国人民的勤劳和智慧, 在科技、工业等各个方面取得了瞩目的成就, 已经在赶超强国的征途中。

一般用“自动化时代”作为第三次工业革命的标签, 其主要的原因在于美国应用数学家诺伯特·维纳 (Norbert Wiener) 创立了控制论, 同时电子计算机的发明和广泛使用以及各种“人机控制系统”的形成, 加速了人类社会从机械化、电气化的时代进入另一个更高级的自动化时代。

1947 年 10 月, 美国应用数学家诺伯特·维纳写出划时代的著作《控制论》, 1948 年出版后, 立即风行世界, 其创立的控制论开创了自动控制的先河。控制论是一门以数学为纽带, 将自动调节、通信工程、计算机和计算技术以及生物科学中的神经生理学和病理学等学科在研究中共同关心的问题联系起来而形成的边缘学科。维纳的深刻思想从多方面突破了传统思想的束缚, 有力地促进了现代科学思维方式和当代哲学观念的一系列变革。

第一次工业革命更多的是工匠的经验使然, 发明家成为一个独立的职业, 而第二次工业革命的一个显著特征是科学技术对工业革命的影响逐步加大。与前两次科技 (工业) 革命相比, 第三次科技革命呈现出许多鲜明特点, 比如 (不限于): ①科学进步与技术开发紧密地结合。②科学与技术的结合在生产中得以产业化, 从而对生产力进行改造, 使生产力发生根本性变革。③军事技术率先突破, 而后带动民用技术, 这是第三次科技革命的重要特征。

事实上, 在这个时间段, 同期革命性发展的领域有很多 (也远非工业这一个领域), 因此人们更愿意用第三次科技革命来表征这个时代。

随着科技的进步, 许多原先仅针对工业革命的学科逐步细分, 就引申出许多新的名词, 比如科学革命、技术革命、工业革命和产业革命这几个既相互区别、又相互联系的概念。一般而言, 科学革命是技术革命的理论基础; 技术革命是在人类改造自然过程中关于制造和操

作的系统知识的社会性和根本性的变革；产业革命是由技术革命引起的，是指国民经济的实际产业结构发生了根本变革，致使经济、社会等方面出现了崭新的面貌。不同的学者对这几个概念的理解各有差异，此处不一一罗列。

第三次科技革命总体上以原子能、电子计算机和空间技术的广泛应用为主要标志，是涉及信息技术、新能源技术、新材料技术、生物技术、空间技术和海洋技术等诸多领域的一场信息控制技术革命。这次科技革命不仅极大地推动了人类社会经济、政治、文化领域的变革，还影响了人类生活方式和思维方式，使人类社会生活和人的现代化向更高境界发展。正是从这个意义上讲，第三次科技革命是迄今为止人类历史上规模最大、影响最为深远的一次科技革命，是人类文明史上不容忽视的一个重大事件。

第三次科技革命对经济发展、社会生活、世界经济、国际关系、全球一体化都产生了重要的影响，比如：它引起生产力各要素的变革，使劳动生产率有了显著提高；它使整个经济结构发生了重大变化，不仅加强了产业结构非物质化和生产过程智能化的趋势，而且引起了各国经济布局 and 世界经济结构的变化；电子计算机的发明和广泛使用以及各种“人机控制系统”的形成，预示着人类社会将从机械化、电气化的时代进入另一个更高级的自动化时代；空间技术和海洋技术的发展标志着人类社会已从被束缚于地球表面的“地球居民”时代进入一个更为辽阔的陆海空立体新时期；基因重组技术、结构化学和分子工程学的进展使人类获得了主动创造新生物和新生命的创造力，标志着人类正在由“必然王国”一步步走向“自由王国”。

(4) 工业 4.0：第四次工业革命

我们现在正处在第四次工业革命的年代。如果说前三次工业革命依次实现了机械化、电气化、自动化的话，那么工业 4.0 的主要特征和目标将是智能化。（德国所谓的）工业 4.0 是指利用信息物理系统（CPS）将生产中的供应、制造、销售等信息数据化、智慧化，最后达到快速、有效、个性化的产品供应。工业 4.0 可以用“1-2-3-4”来描述：

1) 1 个系统。1 个系统指的就是 CPS，CPS 指的是一个综合计算、网络和物理环境的多维复杂系统，通过计算、通信和控制技术的有机融合与深度协作，实现大型工程系统的实时感知、动态控制和信息服务。

本质上说，CPS 是一个具有控制属性的网络，但它又有别于传统的控制系统。传统的控制系统通常是封闭的，即便其中一些工控应用网络也具有联网和通信的功能，但其通信的功能比较弱，往往是通过工业控制总线实现的，且该网络内部各个独立的子系统或者说设备难以通过开放总线或者互联网进行互联。而 CPS 强调的是分布式应用系统中的物理设备互联互通，如果说，如同互联网扁平化和透明化了人与人的交互，CPS 则将所有生产设备、相关的人融合在一个复杂的网络中，从而扁平化和透明化人与物理世界的交互。正如积丰院士认为：CPS 的意义在于将物理设备联网，特别是连接到互联网上，使得物理设备具有计算、通信、精确控制、远程协调和自治五大功能。

2) 2个主题。2个主题指的是智能工厂(Smart Factory)和智能生产(Smart Manufacturing),前者聚焦在生产执行层面,主要内容涉及智能化生产系统及生产流程、网络化分布式生产设施的实现;后者聚焦在企业的运营、研发、管理等宏观层面,主要内容涉及整个企业的生产物流管理、人机互动以及3D技术在工业生产过程中的应用等。

工业4.0战略旨在通过充分利用信息通信技术和CPS相结合的手段,将制造业向智能化转型,目标是建立一个高度灵活的个性化和数字化的产品与服务的生产模式。在这种模式中,传统的行业界限将消失,并会产生各种新的活动领域和合作形式,传统的产业链分工或将被重组并创造出新价值。工业4.0包括智能工厂和智能生产两大主题,鉴于物流在整个价值链中的重要作用,也有学者将智能物流纳入工业4.0的第三大主题,所谓智能物流指的是通过互联网、物联网、物流网,整合物流资源,充分发挥现有物流资源供应方的能力(包括工作效率),而需求方,则能够快速获得服务匹配,得到物流支持。就智能物流的定义来看,从逻辑上可以归属为智能生产的一部分,此处不再赘述。

3) 3个整合。3个整合指的是:在制造企业内部实现网络化制造系统的垂直整合、在不同企业之间实现价值网络的水平整合、在产品整个生命周期内的价值链上实现端到端在工程上的数字整合。

①所谓网络化制造系统的垂直整合(vertical integration and networked manufacturing system),指的是为了提供一种端到端的解决方案,将各种企业内部不同层面的IT系统集成在一起,比如生产环节上的集成(如研发设计内部信息集成)、跨环节的集成(如研发设计与制造环节的集成)和产品全生命周期的集成(如产品研发、设计、计划、工艺到生产、服务的全生命周期的信息集成)。

对于垂直整合而言,需要解决的关键问题是:如何应用CPS系统创建灵活且可重新组合的制造系统?垂直集成是在工厂进行的,在将来的智能工厂里,制造业拓扑结构不会固定并被先期限定,取而代之的是根据个性化需求定制一组IT结构化模块,根据不同情况下产品生产的需要,自动搭建出特定的结构(包括模型、数据、通信和算法等所有响应相关需求的目标架构和解决方案),即通过有效的资源配置(如机器、工作、物流)和其间相互作用关系(如原料周转)完成制造业系统组合。

②所谓价值网络的水平整合(horizontal integration through value network),指的是将各种使用不同制造阶段和商业计划的IT系统集成在一起。这其中既包括一个公司内部的材料、能源和信息的配置(如原材料物流、生产过程、产品外出物流、市场营销),也包括不同公司间的配置(价值网络),这种集成的目标是提供端到端的解决方案。

对于水平整合而言,需要解决的关键问题是:为什么企业通过使用CPS系统,可以将其新商业策略、新价值网络和新商业模式得到持续的支持和实施?显然,这涉及不同单位的合作形式和商业模式以及本单位的可持续发展、商业秘密保护、标准化策略和中长期人员培训

和管理等。

③所谓端到端在工程 (end-to-end engineering across the entire value chain) 上的数字整合,指的是围绕产品全生命周期的价值链创造,通过价值链上不同企业资源的整合,实现从产品设计、生产制造、物流配送、使用维护到产品全生命周期的管理和服务。它以产品价值链创造集成供应商(一级、二级、三级……)、制造商(研发、设计、加工、配送)、分销商(一级、二级、三级……)以及客户信息流、物流和资金流,在为客户提供更有价值的产品和服务的同时,重构产业链各环节的价值体系。

对于端到端在工程上的数字整合而言,需要解决的关键问题是:如何应用 CPS 系统实现包括工程流程在内的端到端的商业过程?从产品开发到制造工程、产品生产和服务,应装备恰当的 IT 系统,为整个价值链提供端对端支撑。一个跨越不同技术学科的、全面的系统工程方法是必需的。

本质上,德国政府提出的“工业 4.0”是出于国家战略实施的双重策略,一方面在制造业中装备 CPS 系统,保持其在全球市场的领导地位;另一方面推广 CPS 技术及产品,建立和培养新的主导市场,进而达到增强德国制造业的目的。本质上,不会对相关行业构成技术层面或与信息技术相关的挑战。相反,技术的发展和进步会推动更新的商业模式和企业的应用模式,这是有助于推动“工业 4.0”成功实施的,当然这更需要所有利益相关者多边共同努力。

在制造领域,全球竞争愈演愈烈,德国不是唯一已经认识到要在制造业中引入物联网和服务的国家,美国、中国、印度等很多国家都已经部署类似的战略计划,尝试通过一系列的机制和财政措施来推动它的发展。

2015 年 7 月 4 日,国务院发布的《关于积极推进“互联网+”行动的指导意见》在“互联网+制造业”的指导意见是:推动互联网与制造业融合,提升制造业数字化、网络化、智能化水平,加强产业链协作,发展基于互联网的协同制造模式。在重点领域推进智能制造、大规模个性化定制、网络化协同制造和服务性制造,打造一批网络化协同制造公共服务平台,加快形成制造业网络化产业生态体系。战略意图及技术思路如上,此处不再赘述。

13.4.4 应用提示

在“互联网+”成为国家战略、行业趋势和潮流的同时,把工业互联网以“三驾马车”之一的角色纳入国家战略,可能的原因在于(不限于):

- 1) 作为一种应用模式,工业互联网在重组和优化了产业链结构、倒逼传统产业转型升级的同时,也为中小企业提供了服务创新的机会。
- 2) 作为一种商业模式,工业互联网在透明和平衡了产供需关系、提升企业弹性制造能力的同时,也为绿色制造提供了资源优化的保障。
- 3) 作为一种富民之径,工业互联网在延展和丰富了消费品市场、拓宽普通百姓消费空间的同时,也为全民创业铺垫了创客空间的支撑。

定制化、柔性化、最优化、自动化、可视化、低碳化是工业4.0的几个要素特征:

1) 定制化指的是随着社会发展,社会化社交网络媒体的兴起,满足客户日益增长的个性化定制服务需求和产品需求,将成为企业的核心竞争力。因此,满足客户个性化需求以及为客户提供差异化的服务和产品等是企业制造的驱动要素。

2) 柔性化指的是实现对小批量多品种生产的支持,甚至对个性化需求的满足。因此,动态的业务与系统工程流程的设计、更多地满足客户个性化的需求、通过新的服务创造价值机会等是企业制造的驱动要素。

3) 最优化指的是整体优化资源的配置,包含使用合适配比的设备、人力和物料达到在合适的时间、合适的地点、用合适的资源高效生产合适数量的合适产品。因此,动态优化的决策、设备产出率和效率、动态的业务与系统工程流程的设计等是企业制造的驱动要素。

4) 自动化指的是运用物联网、M2M信息物理网络等先进技术实现信息与设备间的互动闭环,实现生产线投料、设备控制及数据采集质量控制的自动化。因此,优化的决策、动态的业务与系统工程流程的设计、设备产出率和效率等是企业制造的驱动要素。

5) 可视化指的是实现从“消费者需求→研发设计→生产制造→物流运输→消费者”的全流程的可视化。因此,生产要素信息的全流程跟踪是企业制造的驱动要素。

6) 低碳化指的是从产品设计研发、制造、使用到报废整个生命周期中减少环境污染、节约资源和能源,使资源利用率最高,能源消耗最低。因此应对地球环境变化、应对人口结构变化、绿色制造等是企业制造的驱动要素。

当然,满足上述特征需要多边力量的共同努力,就计算机科学与技术侧而言,针对工业互联网涉及的各个环节提供技术服务、进行技术攻关从而助力细分目标的达成,应当是每一位计算机工作者的使命,也是机遇。作为一种应用示例,本文尝试罗列(但不局限于)一些大数据(技术)在工业互联网领域的若干细分应用场景,下面具体介绍面向产品选型及研发的大数据分析和面向供应商立体画像的大数据分析。

(1) 面向产品选型及研发的大数据分析

作为一个制造型企业的产品设计部门,源于对现有产品的质量改造或新型产品的研发等需求,每天面对的问题包括(不限于):做什么?为什么做?如何做?上述这些问题的解答需要以下信息的汇聚,包括:①哪些地区的哪些人需要什么样的产品?②这个产品市场容量、市场分布及竞品布局怎样以及是否值得做?③如果做的话需要什么样的质量属性?④现有产品的改进维度有哪些?

为了有效回答上述问题,有必要建立一个大数据平台,这个数据平台中不仅要包含企业内部的营销数据,还要集成更多的外部数据,如行业监管门户、政府门户(发改委、科技部、国务院、公、检、法等)、社交媒体、新闻网站、企业门户、电商网站(阿里巴巴、亚马逊等)、社交媒体、新闻网站、论坛,当然这些数据的采集一般需要商务的支撑,比如向合作单位购买。在上述数据的基础上,通过构建数据分析平台(一般包含数据服务和计算服务),实现市场分析系统、产品分析系统、竞品分析系统、竞争对手分析系统、舆情分析系统等一些

应用,综合这些应用达到准确定位产品和市场、认清对手、认清自己的目的,从而为产品选型及研发提供信息服务。

一般而言,市场分析系统需要完成的功能包括行业评估和产业链评估,前者通过对产品的目标市场份额、用户需求规模、行业年报等的有效分析,为企业战略规划提供辅助决策;后者通过对产品涉及上下游产业链的目标市场份额、用户需求规模、行业年报等的有效分析,为企业战略规划提供辅助决策。

产品分析系统需要完成的功能包括客户群分析、营销渠道分析、价值评估、产品选型评估。客户群分析通过对产品涉及的用户群进行有效分析,达到对客户群规模、年龄分布、地域分布等精准把握的同时,建立“产品-用户”关联模型;营销渠道分析通过对销售渠道及销售数据的有效分析,建立“产品-渠道”关联模型,结合既有渠道优势,为企业渠道规划提供数据支撑;价值评估通过对产品的目标市场、上下游产业链的成熟度、生产成本以及网络舆情(尤其是需求和情感)的有效分析,评估产品价值;产品选型评估通过“产品-用户-渠道-价格”模型(两两)的构建,为企业产品选型提供辅助决策支撑。

竞品分析系统需要完成的功能包括售前营销分析、销售数据分析、售后舆情分析等。售前营销分析通过对竞品的销售广告、套餐、渠道的有效分析,为企业的战略规划提供辅助支撑;销售数据分析需要对地域分布、价格区间、人群分布进行有效的分析,其中,地域分布分析用于了解竞品销售地域分布的同时建立“产品-地域”模型,辅助企业进行销售渠道布点决策,价格区间分析用于了解竞品销售地域分布的同时建立“产品-价格”模型,辅助企业进行销售渠道布点决策,人群分布分析用于了解竞品客户群分布的同时建立“产品-用户”关系模型,辅助企业进行精准营销;售后舆情分析通过对竞品售后的舆情数据(特别是负面信息)进行有效分析,为企业的产品质量改进和产品设计提供辅助决策。

对手分析系统需要完成的功能包括自然属性分析、主营产品分析、新闻事件分析、企业舆情分析、营销事件分析等。自然属性分析通过对对手的营销规模、财务年报、组织架构、市场渠道、上下游供应商等的有效分析,为企业战略决策提供数据支撑;主营产品分析通过对对手的主营产品及其销售区域、销售量及主营产品关系的有效分析,结合“产品分析”的结果,达到对对手产品的全方位把握;新闻事件分析通过对涉及对手的新闻事件(含涉及的人和组织)以及网络反馈事件的有效分析,为企业战略决策提供辅助决策;企业舆情分析通过对涉及对手的网络舆情的有效分析,达到对对手互联网认可度的全面把握,为企业战略决策提供数据支撑;营销事件分析通过对对手的营销事件、活动内容、相关事件和人的有效分析,为企业营销事件规划提供辅助支撑。

舆情分析系统需要完成的功能包括舆情简报、热点舆情、情绪舆情、风险预警和热词发现等。舆情简报以某周期(日、周、月)统计、归纳所关注舆情摘要(含去重),为企业舆情监管提供数据支撑;热点舆情通过对网络数据进行实时语义分词分析,达到对热点事件、敏感词或事(可配置)实时预警的目的;情绪舆情通过对网络数据进行实时情感语义分析,达到对负面情感(可配置,可动态更新,形成企业情感数据库)实时预警的目的;风险预警通

通过对风险因子关键字（企业设定或参照行业标准）的实时发现（热点舆情），达到对风险组合因子的建模和主动预警的目的；热词发现通过对网络数据的统计分析，发现新闻（事件）驱动的新词，从而达到对敏感词数据库及情感数据库等的自动更新的目的。

对于产品设计与研发部门而言，通过上述大数据分析可以获得有力支撑，比如：根据产品分析系统了解哪些用户需要哪些产品；根据市场分析系统可以得到哪些产品值得做以及需求规格如何；通过竞品分析系统，可以得到用户希望产品应当具备的质量属性；根据对手分析系统可以了解此产品的市场前景和分布；根据舆情分析系统可以了解现有用户对企业的改进需求……

（2）面向供应商立体画像的大数据分析

作为一个制造型企业的采购部门，每天面对的问题包括（不限于）：哪些供应商备选？为什么要选择这些供应商？供应商（供应链）是否有风险？上述这些问题的解答需要以下信息的汇聚，包括：①当前供应商是否（还）值得合作？②还有哪些备选的供应商？③现有供应商是否存在运营风险？大数据分析能够有效回答上述问题。大数据平台的建设思路如上所述，此处不再赘述。针对供应链立体画像的分析目标，在上述大数据分析平台的基础上构建健康度评估系统、价值度评估系统、新供应商发现系统等，借此实现对供应商的立体画像以实现高效的辅助决策。

健康度评估系统需要包括的功能有风险策略配置、风险主动预警、舆情简报摘要等。其中风险策略配置以企业价值评估体系为标准配置关注企业名录、数据源清单、风险因子数据库、敏感词数据库；风险主动预警以风险策略配置为中心，结合大数据平台及舆情分析系统，达到对敏感事件、新闻等的实时推送；舆情简报摘要结合舆情分析系统，以某周期（日、周、月）摘要推送所关注企业舆情信息等。

价值度评估系统的主要目标是对供应商的实际价值做出客观准确的评估，能够动态地定制评估模块的组合并对模块赋予合适的权值，形成可定制的评估模型，需要包括的功能有收入能力评估、扩展能力评估、网络信誉评估。其中，收入能力评估从企业体系的内部成本折算和供应商主营产品（产能）综合评估供应商的生产能力、行业资信等；扩展能力评估通过对供应商的主营产品（系列）的组合分析，评估其在企业体系的价值度（单一产品的供应商或系列产品的供应商）；网络信誉评估结合大数据平台及舆情分析系统，对供应商的网络信誉进行综合评估等。

新供应商发现系统需要的功能包括候选实体发掘和候选实体推荐等。其中候选实体发掘从主营、对手、行业等多个层次进行发掘，主营驱动的发掘指的是以供应商主营产品（需求）以及供应商相似竞争对手为中心进行全网检索，对手驱动的发掘指的是通过对竞争对手的上下游供应商进行有效分析，行业驱动的发掘指的是通过行业新闻、年报进行有效分析；候选实体推荐从健康度、价值度、地域、网络（行业）口碑等多维角度对候选实体进行打分，给出候选实体的排序。

对于采购部门而言,通过上述大数据分析可以获得有力支撑,比如:根据供应商价值评估系统,得到当前供应商是否值得合作的辅助意见;如果需要更换供应商,新供应商发现系统可提供候选供应商;而供应商健康度评估系统会及时提示某供应商可能存在问题(风险)及原因,以供人工复核。

13.5 互联网金融

13.5.1 基本概念

金融是指货币的发行、流通和回笼、贷款的发放和收回、存款的存入和提取、汇兑的往来等经济活动,其中三个主要的活动是投资、融资和支付(其实这三个也是三大基本金融需求)。投资是资金盈余者(提供者)想用钱生钱(比如储蓄);融资是资金短缺者(需求者)想用钱买钱(比如贷款);相对于投融资活动,支付相对比较独立,指的是发生在购买者和销售者之间的金融交换过程。具体如表13-3所示。

表 13-3 金融活动

	投资活动	融资活动	支付活动
涉及主体	资金盈余者	资金短缺者	所有参与者
目的	钱生钱	钱买钱	资金流动
产品示例	储蓄、基金理财	银行贷款、民间借贷	钱-物交换
	余额宝、理财通	阿里小贷、拍拍贷	支付宝、微信支付

在男耕女织的农业经济时代,支付是主要的金融活动,而且往往是基于纸币的,流通范围有限,且资源配置手段单一,宏观调控不易进行,这样的金融时代往往被冠以“金融1.0”的标签。

随着时代的进步和工业经济的发展,传统的纸币金融已经远远不能跟上商务活动的范围和广度不断扩大的节奏,于是很自然地过渡到被冠以“金融2.0”的数字金融时代。这个时代的一个典型特点是各种银行卡的发行和应用,通过银行卡实现电子支付完成各种交易活动,具有保存成本低、流通费用低、标准化成本低、使用成本低等诸多优势。同时,也由于数字中心的集中管控,国家宏观调控和配置也具有较大的便利。

当前,我们进入了被冠以“金融3.0”的互联网金融时代,互联网金融(Internet Finance或Online Financial)是互联网技术、移动通信技术实现资金融通、支付和信息中介等业务的新兴金融模式,既不同于商业银行间接融资,也不同于资本市场直接融资的融资模式。

2014年3月5日,李克强总理在政府工作报告中首次提及对互联网金融的有效监管:“……促进互联网金融健康发展,完善金融监管协调机制,密切监测跨境资本流通,守住不发生系统性和区域性金融风险的底线。让金融成为一池活水,更好地浇灌小微企业、‘三农’等实体经济之树……”

2015年3月5日,李克强总理在政府工作报告中提及“互联网+”的国家战略时,把互联网金融作为重点发力的“三驾马车”之一。

2015年7月4日,国务院公开发布的《关于积极推进“互联网+”行动的指导意见》(国发〔2015〕40号)再次把互联网金融作为重点行动,并冠之以“互联网+普惠金融”的战略目标:“……促进互联网金融健康发展,全面提升互联网金融服务能力和普惠水平,鼓励互联网与银行、证券、保险、基金的融合创新,为大众提供丰富、安全、便捷的金融产品和服务,更好满足不同层次实体经济的投融资需求,培育一批具有行业影响力的互联网金融创新企业……”

互联网金融的本质还是金融,因此所有的互联网金融产品都是围绕着金融的主要活动进行的,不过鉴于互联网及金融的特点,又衍生出很多的产品业务模式,如表13-4所示。

表 13-4 互联网金融典型业务模式

投资活动		融资活动		支付活动		风险管理		其他	
网络银行				第三方支付		网络保险		金融产品搜索引擎	
P2P 借贷平台									
众筹平台									
网络资产交易平台									
网络基金		网络微贷		网络征信		其他			
网络证券									
网络理财									
财富管理									

以下几节对上述的业务模式进行简单的介绍。

13.5.2 面向投融资的互联网金融

说到底,投融资活动的本质是资金盈余者将资金转移给资金短缺者并因此获利,投融资活动涉及的主体包括资金盈余者、资金短缺者,以及为了更高效、安全、便捷地进行投融资活动而出现的金融机构和第三方服务。这四大主体按照不同的投融资规则,形成了以下几个不同的投融资活动:

(1) A 模式:资金盈余者→资金短缺者

此模式是最简单的一种投融资模式,即资金盈余者直接将资金转移给资金短缺者,比如个人与个人之间的民间借贷,个人A将资金借给个人B获得收益(利息或者其他等价交换物,情感的维系也算是一种收益)。相对而言,这种投融资模式无论是投融资效率还是盈余资金的使用效率都是最低的,因为不是所有的人都能够在自己熟悉的朋友圈中找到愿意借款的资金盈余者,而资金盈余者也不是那么容易找到愿意且自己信得过的人进行投资。金融机构的出现很好地弥补了这方面的不足。

个人与个人之间的民间借贷往往是基于个人与个人之间的熟悉关系以及因为这种熟悉关系而由个人自己建立起来的评估信用体系。或者说这种熟人互评的信用是维系这种投融资的重要保障（当然，这期间或许也会有抵押物或者借款协议等），应该说，这种信用评估体系是在一个长期交往的基础上建立起来的，建立的成本比较大比较重，也更加可信。但是这种信用评估几乎仅适用于借贷双方而无法更好地加以复用，事实上是一种浪费。2015年，有一款基于熟人社交关系的互联网金融产品“自信”（暂且不论该产品是否成功，产品是否成功还与商业运作、市场环境、系统运维等诸多因素相关）则是充分考虑到这种熟人之间的互评建立起来的每个人的信用评估，将所有的这些信用评估数据收集起来，建立每个人的信用评估体系，并用于个人与个人之间的P2P借贷，算是针对这个场景，利用互联网思维进行的创新应用。

（2）B模式：资金盈余者→金融机构→资金短缺者

本质上，这种模式是上述A模式的变种，金融机构承担着资金盈余者和资金短缺者中介的角色：一方面，资金盈余者将资金转移给金融机构（金融机构充当资金短缺者的角色）；另一方面，金融机构将资金转移给资金短缺者（金融机构充当资金盈余者的角色）。显然，专门的金融机构的介入会让投融资的效率和资金使用率大幅提高，但是这三方承担的风险和风险承受能力是不一样的，比如资金盈余者会担心资金转移给金融机构是否安全，是否会获得预期的收益；资金短缺者虽然能获得资金融资，但是否能够如约归还给金融机构；而金融机构所承担的风险最大，它一方面要有足够多的渠道能够向更多的资金盈余者吸储，另外一方面要有足够多的渠道向更多资信良好的资金短缺者放贷，更重要的是它还要承担资金发放出去以后的风险（是否能够如约收回资金短缺者的款项，以及是否能够从前后两端的利息差中获益）。因此，风险控制（通常简称为风控）是每个金融机构的重中之重。

具体而言，金融机构面对的风险还可以进一步分为市场风险、信用风险和操作风险，市场风险指的是因市场波动使得投资者不能获得预期收益的风险；信用风险指的是合同的一方不履行义务的可能性，包括贷款、掉期、期权及在结算过程中的交易对手违约带来损失的风险；操作风险指的是因交易或管理系统操作不当导致损失的风险，包括因公司内部失控而产生的风险。

在“互联网+”驱动下的金融产品和服务的互联网改造往往包括下面三种类型：

1) 金融机构互联网化。由金融机构参与的大部分金融产品（包括存款、贷款、基金、信托、保险等）都需要传统金融机构的牌照去设计、打包和生产，这意味着传统金融机构具有典型的江湖垄断色彩，而为了迎接互联网时代的各类压力和挑战，传统金融机构务必对既有产品进行互联网改造。针对这种层面的“互联网+金融”，人们更愿意用金融互联网（而不是互联网金融）来表示。

2) 互联网公司金融机构化。这种互联网金融的模式是：传统互联网公司或电商企业通过获得牌照或入股具有牌照的金融机构开展相关金融业务。比如阿里巴巴和腾讯参股的民营银

行陆续获批,阿里巴巴、苏宁、百度、京东陆续设立小贷公司,从事传统金融机构的信贷业务。

3) 金融机构与互联网公司合作开发。这种互联网金融的模式是互联网企业和传统金融机构基于优势互补的理念,展开强强联合,共同开发互联网金融产品。比如阿里巴巴和天弘基金合作开发的余额宝产品;腾讯、阿里、平安共同开发互联网保险产品。

(3) C 模式:资金盈余者→第三方服务→资金短缺者

在最基本的投融资流程的基础上,由于信息不透明,需要投融资双方以外的第三方介入,以促成供求信息配对,完成投融资活动,这就引发了此模式的兴起。此处的第三方与传统金融机构的典型区别在于:此处的第三方不会吸收资金和出借资金,而是仅作为撮合资金盈余者和资金短缺者的中介,仅提供信息服务或者渠道服务,也有人用金融电商化来描述这种金融产品的应用模式。

普通消费者通过电子商务(平台)可以找到合适的产品,或者说厂家通过电子商务(平台)可以找到合适的买家,电子商务(平台)提供给买卖双方的本质上是一种信息沟通平台或者信息服务。互联网金融的本质是金融,所有金融活动的目标都可归结为开发一系列产品满足投融资需求。金融电商化就是将金融产品以互联网为平台、以电子商务为渠道建立“产品—个人(组织)”的桥梁,从而更好地完成投融资活动。如前所述,根据电子商务参与的角色不同,可以将电子商务的应用模式分为 B2C、B2B、C2C 等,在金融电商化的场景下,这里的 B 对应的是金融机构,而 C 对应的是非金融组织或者个人。或许,类比于电商的其他应用模式,金融电商也会创新出其他类似的电商模式。

一般用于投融资活动的互联网金融产品业务模式有如下几种(不限于):

1) P2P 借贷指的是个人对个人的投融资活动(如人人贷、拍拍贷、宜信、陆金所),起源于英国、随后发展到美国和其他国家,其典型模式为网络信贷公司提供平台,由借贷双方自由竞价以撮合投融资活动的成交。根据网贷平台在投融资活动中的角色不同,又可具体分为以下几种:

①网贷平台仅为借贷双方提供信息流通交互、信息价值认定和其他促成交易完成的服务,不实质参与到借贷利益链条之中。借贷双方直接发生债权债务关系,网贷平台则依靠向借贷双方收取一定的手续费维持运营。本质上是上述的 C 模式投融资模式,即资金盈余者→第三方服务→资金短缺者。

②网贷平台自身或者通过和担保公司、小贷公司、理财公司、融资租赁公司、证券公司、银行合作,充当第三方结构做风控,网贷平台把做好风控的项目放在平台上和投资者对接,从而化解平台做风控不专业,风控水平较差的难题。本质上是上述的 C 模式投融资模式,即资金盈余者→第三方服务→资金短缺者。

③网贷平台利用债权转让模式,充当连接借款方和贷款方的中间金融机构,从而可以主动批量地量化开展业务,而不是被动等待各自的匹配,从而实现规模的快速扩展。本质上是

上述的B模式投融资模式,即资金盈余者→金融机构→资金短缺者。

2) 众筹平台:是一种大众筹资或群众筹资的投融资模式(如点名时间、淘梦网、众意网),具体而言,是通过网络平台连接起发起人(有创造能力但缺乏资金的人)和支持者(对筹资者的故事和回报感兴趣的并且有能力支持的人),根据支持者获益回报的方式,又分为债券融资、股权融资、整体转让等多种,一般需要(第三方)金融机构充当众筹平台的资金监管。

3) 网络资产交易平台:通过网络交易平台为个人投资者建设快捷财富增值通道,实现个人与各银行间的一站式互联网金融资产交易。

4) 网络微贷:是指互联网企业通过其控制的小额贷款公司,利用互联网向客户提供的小额贷款金融服务。一般而言,投融资相关的一切认证、记账、清算和交割等流程均通过网络完成,借贷双方足不出户即可实现借贷目的,而且一般额度都不高,无抵押,纯属信用借贷。随着中国的金融管制逐步放开,在中国巨大的人口基数、日渐旺盛的融资需求、落后的传统银行服务状况下,这种网络借贷新型金融业务有望在中国得到爆发式增长。

5) 其他类投融资产品:基于互联网思维,将互联网作为渠道和服务平台,将一些传统的金融产品互联网化,实现人的各类投融资活动,比如网络基金、网络证券、网络理财、网络管理。

13.5.3 面向支付的互联网金融

支付是金融活动中最原始也是最基础的活动,也是撬动互联网金融的最重要的驱动源。从有电子支付开始(纸币金融时代的那种“一手交钱一手交货”模式此处不议),支付也经历了若干变迁,粗分为3类:

1) 支付1.0时代:以网上银行为代表,支付作为金融工具完成交易。网上银行包含两层含义:一个是机构概念,指通过信息网络开办业务的银行;另一个是业务概念,指银行通过信息网络提供的金融服务,包括传统银行业务和因信息技术应用带来的新兴业务(如网上投资、网上购物、个人理财、企业银行)。在日常生活和工作中,我们提及网上银行,更多是第二层次的概念。网上银行系统是银行业务服务的延伸,客户可以通过互联网方便地使用商业银行核心业务服务。如何保证网上银行交易系统的安全关系到银行内部整个金融网的安全,这是网上银行建设中至关重要的问题,也是银行保证客户资金安全的最根本的考虑。网上银行系统一般都采用加密传输交易信息的措施,使用最广泛的是安全套接层(Secure Socket Layer, SSL)协议和安全电子交易(Secure Electronic Transaction, SET)协议。

SSL协议是由Netscape公司推出的一种安全通信协议(作用在传输层),其首要目的是在两个通信主体间提供秘密而可靠的连接(用户登录并通过身份认证之后,用户和服务方之间在网络上传输的所有数据全部用会话密钥加密,直到用户退出系统为止)。而且每次会话所使用的加密密钥都是随机产生的,这样,攻击者就不可能从网络上的数据流中得到任何有用的信息。同时,引入了数字证书对传输数据进行签名,一旦数据被篡改,则必然与数字签名不

符。SSL 协议的加密密钥长度与其加密强度有直接关系,一般是 40~128 位。本质上,SSL 是一个保证计算机通信安全的协议,SSL 协议运行的基础是商家对客户信息保密的承诺而进行的一种商家对用户的认证。按照 SSL 协议,客户的购买信息首先发往商家,商家再将信息转发银行,银行验证客户信息的合法性后,通知商家付款成功,商家再通知客户购买成功,并将商品寄送客户。但在这个过程中缺少客户对商家的认证,在电子商务的开始阶段,由于参与电子商务的公司大都是一些大公司,信誉较高,这个问题没有引起人们的重视。随着电子商务参与的厂商迅速增加,对厂商的认证问题越来越突出。针对这种情况 SET 协议逐步得到重视。

一个基于 SSL 协议的支付涉及的主体包括消费者、商家、银行等,所有的主体均由认证中心进行身份的确定,确保交易双方的合法身份,支付流程如图 13-3 所示。

①消费者和商家经过充分洽谈达成交易意向后,通过 SSL 将订单及请求支付信息发往商家服务器,如图中步骤①。

②商家将消费者发来的订单及请求支付信息通过 SSL 经由支付网关发往银行服务器,如图中步骤②。

③银行审核后将确认信息经由支付网关通过 SSL 发往商家服务器,如图中步骤③。

④商家服务器将确认信息通过 SSL 发送消费者客户端,如图中步骤④。

SET 协议是由 VISA 和 MasterCard 两大信用卡公司于 1997 年 5 月联合推出的规范(作用在应用层),SET 主要是为了解决用户、商家和银行之间通过信用卡支付的交易问题而设计的,以保证支付信息的机密、支付过程的完整、商户及持卡人的合法身份以及可操作性。SET 在保留对客户信用卡认证的前提下,又增加了对商家身份的认证,这对于需要支付货币的交易来讲是至关重要的。

一个基于 SET 协议的支付流程如图 13-4 所示。

一个基于 SET 协议的支付涉及的主体包括消费者、商家、银行等,所有的主体均由认证中心进行身份的确定,确保交易双方的合法身份,支付流程如下:

①消费者和商家进行充分沟通和协商达成交易意向,如图中步骤①。

②消费者填写订单并将完整的订单和请求付款的指令发给商家(所有信息由持卡人进行数字签名,商家看不到消费者账号信息),如图中步骤②。

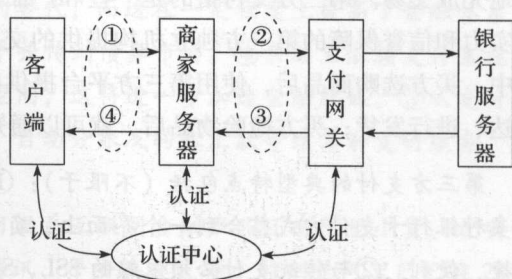


图 13-3 基于 SSL 协议的支付流程

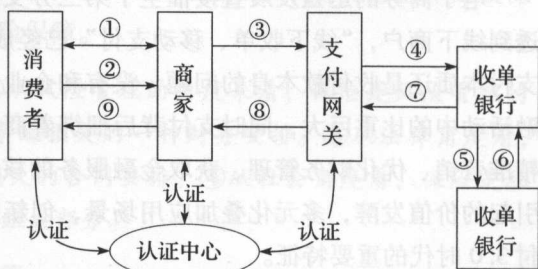


图 13-4 基于 SET 协议的支付流程

③商家收到订单后,向持卡人的金融机构请求支付认可,通过支付网关到收单银行,然后再到发卡银行,如图中步骤③→步骤④→步骤⑤。

④发卡银行审核后将批准交易信息发送给收单银行,再通过支付网关,发送到商家,如图中步骤⑥→步骤⑦→步骤⑧。

⑤商家发送订单确认信息给消费者,如图中步骤⑨。

事实上,为了保障网上交易的安全,还有若干安全保障技术,此处不再赘述。

2) 支付 2.0 时代:以第三方支付为代表,支付作为交易中介,协助双方高效、透明、安全地完成交易。第三方支付指的是和一些和产品所在国家以及国内外各大银行签约、并具备一定实力和信誉保障的第三方独立机构提供的交易支付平台。在通过第三方支付平台进行的交易中,买方选购商品后,使用第三方平台提供的账户进行货款支付,由第三方通知卖家货款到达、进行发货;买方检验物品后,就可以通知付款给卖家,第三方再将款项转至卖家账户。

第三方支付的典型特点包括(不限于):①第三方支付平台提供一系列的应用接口程序,将多种银行卡支付方式整合到一个界面上,负责交易结算中与银行的对接,使网上购物更加快捷、便利。②与传统支付必须依赖的 SSL、SET 等支付协议相比,利用第三方支付平台进行支付操作更加简单而易于接受。③第三方支付平台本身一般依附于大型的门户网站,且以其合作的银行的信用作为依托,因此第三方支付平台能够较好地突破网上交易中的信用问题,有利于推动电子商务的快速发展。

第三方支付的诞生源于非现金交易的需求,第三方支付作为中介机构介入其中,承担便利支付和信用中介(前文提到这是支付宝首创)等职能,促成交易发生。2010年,中国人民银行制定了《非金融机构支付服务管理办法》,并于次年开始发放第三方支付牌照,至此,第三方支付正式纳入央行金融监管体系。所谓第三方支付牌照,指的是支付业务许可证,只有获此许可证,才能从事第三方支付业务。截至2015年3月30日,中国人民银行共发放270张牌照。特别值得注意的是:根据具体从事的业务差异,第三方支付牌照又细分为互联网支付、银行卡收单、预付卡发行和受理、固定及移动电话支付和数字电视支付牌照等。

3) 支付 3.0 时代:支付不仅是一种工具或交易中介,还可以作为渠道,叠加更多的金融服务、营销服务和征信服务等,从而成为互联网金融的核心。

电子商务的迅猛发展直接催生了第三方支付(支付 2.0 时代的标志)的兴起,并很快渗透到线下商户,“线下收单、移动支付”已经成为一个普适的应用模式。但是“支付 2.0”的支付本质还是收付款本身的问题,没有和企业业务相关联。一个事实是:支付本身占整个金融活动中的比重巨大,同时支付背后捆绑着商户最真实的资金流和信息流,这都是企业实现精准营销、优化财务管理、获取金融服务的核心依托。因此,如何有效利用支付,扩展支付引起的价值发酵,多元化叠加应用场景,创新以支付为翘板的互联网金融应用模式,成为支付 3.0 时代的重要特征。

2013年,互联网金融产品余额宝的横空出世,触动了几乎所有传统金融大鳄的神经。本

质上,余额宝是一个货币基金,按照传统的金融视角,货币基金好像并不会对传统金融机构(比如银行)有太大的冲击(事实上,货币基金对银行而言,也有获利)。但是余额宝是嫁接在支付宝(第三方支付平台的航母)基础上的,这个货币基金对传统金融的冲击就不再一般:①分流部分客户储蓄,本来理财产品已经对银行吸储产生了较大的压力,而余额宝背后八亿支付宝注册用户则更是加剧了银行的这种压力。②虽然还未对银行的基金理财产品形成直接冲击,但是银行中间业务收入受到直接冲击,比如余额宝直接影响银行基金代销渠道。③余额宝形成的示范效应,加速了市场上各种理财服务的涌现,在助推互联网金融市场发展、加速金融脱媒的同时,形成货币基金对银行存款的替代,而这在一定程度上提高了金融体系的金融成本并大幅拉高了金融体系短期资金率水平并传到债券市场。④余额宝依赖支付宝庞大的客户资源和资金沉淀给了余额宝巨大的发展空间,这将进一步加速金融脱媒。⑤从支付角度看,第三方支付平台已能为客户提供收付款、自动分账及转账汇款等结算和支付服务,逐渐替代传统银行支付业务功能。

除了余额宝外,作为第三方支付撬动起的阿里系互联网金融产品还有阿里小贷、众安保险(国内首个互联网保险公司,由阿里巴巴、腾讯、平安共建)等,特别是阿里巴巴在2015年获得个人征信牌照后立刻推出的芝麻信用评估系统,使得阿里在互联网金融体系内的金融服务范围越来越广,越来越完备。而在阿里系互联网金融不断推出的同时,其他第三方支付平台公司陆续推出的互联网金融产品也层出不穷。作为金融的痛点,支付势必会进一步成为互联网金融产品不断创新的原始驱动力和发展机遇。

13.5.4 其他类型的互联网金融

互联网金融的本质是金融,而相关的金融活动除了上述的投融资和支付之外,还有一些相关需求:

(1) 保险

作为金融体系和社会保障体系的重要支柱,保险是指投保人根据合同约定,向保险人支付保险费,保险人对于合同约定的可能发生的事故造成所导致的财产损失承担赔偿责任,或者当被保险人死亡、伤残、疾病或者达到合同约定的年龄、期限等条件时承担给付保险金责任的商业保险行为。本质上这是与传统的钱物交易一致的金融活动,交易主体是投保人和保险人(法人),而其中的“物”是一种风险保障。

从不同的视角来看保险,有不同的认知:①从风险管理的角度来看,保险是风险管理的一种方法。②从经济角度看,保险是分摊意外事故损失的一种财务安排。③从法律角度看,保险是一种合同行为,是一方同意补偿另一方损失的合同安排。④从社会角度看,保险是社会生产和社会生活的稳定器,是社会保障的重要组成部分。

保险具有经济补偿、资金融通和社会管理功能:

1) 经济补偿功能是基本的功能,也是保险区别于其他行业的最鲜明的特征。



2) 资金融通功能是在经济补偿功能的基础上发展起来的,资金融通的功能是指将形成的保险资金中闲置的部分重新投入到社会再生产过程中,这是保险人为了使保险稳定经营而必须保证保险资金的增值与保值的必然金融活动。

3) 社会管理功能是保险业发展到一定程度并深入社会生活诸多层面之后产生的一项重要功能,是指对整个社会及其各个环节进行调节和控制的过程,目的在于正常发挥各系统、各部门、各环节的功能,从而实现社会关系和谐、整个社会良性运行和有效管理。

互联网保险业务是指保险机构依托互联网和移动通信等技术,通过与自营网络平台、第三方网络平台等订立保险合同,提供保险服务的业务(《互联网保险业务监管暂行办法》,保监发〔2015〕69号),即保险信息咨询、保险计划书设计、投保、交费、核保、承保、保单信息查询、保全变更、续期交费、理赔和给付等保险全过程均通过互联网完成。相比较于传统保险而言,互联网保险具有很多独特的优势,比如:

1) 对于客户而言,互联网保险让客户能自主选择产品,客户可以在线比较多家保险公司的产品,保费透明,保障权益也清晰明了,且各类咨询服务和业务办理也较方便。

2) 对于保险公司而言,保险公司也能从互联网保险中获益多多,比如险种的设计、保险计划的设计和营销等方面的费用减少。

鉴于互联网保险在金融领域的重要地位,为规范互联网保险经营行为,促进互联网保险健康规范发展,2015年7月22日,中国保监会印发了《互联网保险业务监管暂行办法》,标志着我国互联网保险业务监管制度正式出台。暂行办法首先对互联网保险进行了官方定义,并强调不能确保客户服务质量和风险管控的保险产品,保险机构应及时予以调整,同时,互联网保险消费者享有不低于其他业务渠道的投保和理赔等保险服务。此外,在经营条件、经营区域、信息披露、经营规则、监督管理等方面也都提出了明确的要求。

(2) 征信

征信就是专业化的、独立的第三方机构为个人或企业建立信用档案,依法采集、客观记录其信用信息,并依法对外提供信用信息服务的一种活动。征信服务可以为防范信用风险、保障交易安全创造条件,促进形成诚信者受益、失信者受惩戒的社会环境。

电子商务是基于互联网的一个交易行为,如何保障交易的安全、可信是交易得以顺利进行的关键。以支付宝为例,支付宝可以看作是由淘宝作为中介的交易信用担保:买卖双方经过沟通达成交易意向,买家将钱打入支付宝,卖家发货,买家收货(确认)后,卖家从支付宝中获得先前买家支付的费用。因为有支付宝,卖家无须担心货发了却收不到款,买家也无须担心钱支付了却收不到货。支付宝的担保模式事实上使买卖双方都有一个潜意识里的信用认同:买家相信支付宝能够保障他收到货,而卖家也相信支付宝能够保障他收到款。事实上,担保公司大多也承担类似的职能,只不过不是在淘宝的平台上进行而已。

信用本质是一种债权债务关系,即授信者相信受信者具有偿还能力,而同意受信者所做的未来偿还的承诺。但当商品经济高度发达,信用交易的范围日益广泛时,信用交易的一方想要

了解对方的资信状况就会极为困难。此时,了解市场交易主体的资信就成为一种需求,征信活动也应运而生。

《征信业管理条例》是经2012年12月26日国务院第228次常务会议通过且于2013年1月21日中华人民共和国国务院令631号公布的文件。该《条例》分总则、征信机构、征信业务规则、异议和投诉、金融信用信息基础数据库、监督管理、法律责任、附则8章47条,自2013年3月15日起施行。该条例的出台,解决了征信业发展中无法可依的问题,有利于加强对征信市场的管理,规范征信机构、信息提供者和信息使用者的行为,保护信息主体权益;有利于发挥市场机制的作用,推进社会信用体系建设。

按照不同的分类标准,征信大致分为以下几类:

1) 按业务模式可分为企业征信和个人征信两类,前者主要是收集企业信用信息、生产企业信用产品的机构;后者主要是收集个人信用信息、生产个人信用产品的机构。相对而言,企业征信成熟度较高,国外有很多成熟的经验可以参考。目前我国企业征信牌照的发放以备案制为准,个人征信牌照以审核制为准。

2) 按服务对象可分为信贷征信、商业征信、雇佣征信以及其他征信。其中,信贷征信主要服务对象是金融机构,为信贷决策提供支持;商业征信主要服务对象是批发商或零售商,为赊销决策提供支持;雇佣征信主要服务对象是雇主,为雇主用人决策提供支持;另外,还有其他一些征信活动,诸如市场调查,债权处理,动产、不动产鉴定。

3) 按征信范围可分为区域征信、国内征信、跨国征信等。其中,区域征信一般规模较小,只在某一特定区域内提供征信服务;国内征信是目前世界范围内最多的机构形式之一,近年来开设征信机构的国家普遍采取这种形式;跨国征信这几年正在迅速崛起,此类征信之所以能够得以快速发展,主要原因大致有:①西方国家一些老牌征信机构为了拓展自己的业务,采用多种形式向其他国家渗透。②由于世界经济一体化进程的加快,跨国经济实体越来越多,跨国征信业务的需求逐步增强。

从大数据的角度来看,征信本身一定是大数据。以信用体系最全的美国为例,5C1S (Character、Capability、Capital、Condition、Collateral、Stability) 是美国对借款对象进行贷前分析和审查时惯用的评价体系,主要使用以借款者的基本信息、财务状况和过往借贷行为等与借贷对象经济行为直接相关的数据为基础,而这些数据也正是美国各大征信机构主要收集的数据。与美国类似,我国央行征信中心也主要以个人基本数据、金融数据(主要是信贷和信用卡相关数据)、公共数据(包括税务、工商、法院、电信、水电煤气等部门的数据)以及个人信用报告查询记录这四个类别为主。事实上,可以看到,传统征信所使用的数据已经完全具备了大数据所谓的4V特征。而互联网技术及应用的普及,使得互联网中以数据通信为载体的个人信息及行为信息更易被采集,互联网和大数据技术使得可以用来评估的数据维度越来越丰富,如电商的交易数据、社交类数据、网络行为数据,所有上述互联网引发的现象使得征信数据的大数据特征更加明显和突出。

随着征信大数据时代的到来,如何迎接和响应其中的挑战,这是摆在每一个大数据相关工作者面前并需要其关心的问题。

(3) 信息服务

互联网企业发展金融业务的几种模式除了前文提到的第三方支付(如支付宝)、在线信贷(如阿里小贷)、基金销售(如余额宝)外,还有一种应用模式称为流量分发(如融360、好贷网),该模式很好地迎合了金融活动中有关交易双方对称信息的需求。所谓流量分发模式,指的是提供金融产品搜索、搜索引擎推荐和服务平台,为中小企业、小微企业和消费者提供免费、高效、便捷、划算的金融服务,同时通过交易额的某种比例获得服务收益。

流量分发的一般运行模式是:用户在网上输入贷款金额、期限以及用途等关键词,系统就会进行比对和处理,输出一份相应的银行及其他信贷机构的列表(包括银行名称、信贷产品、利率、总利息、月供、放宽时间和贷款总额等信息)。用户进行比较后,可以在线填写申请材料,申请一家或者几家银行的贷款。申请完后,相关银行的信贷经理会与申请人进行电话联系,确认信息,申请人可以再次比较各家银行的产品,之后就可以去分行或支行申请贷款。

流量分发模式相当于介于金融双方的第三方,仅提供信息服务(13.5.2节提及的投融资C模式:资金盈余者→第三方服务→资金短缺者)。其相当于金融领域的百度,能够为用户提供便捷、安全、免费的投融资搜索和推荐服务,同时也为金融机构提供互联网金融创新服务:对于金融机构而言,传统营销和市场的做法成本高、效率低;中介信用低、作假严重;客户数据基础难以支持产品创新和风险控制;而对于有资金需求的用户而言,金融产品纷繁复杂,缺乏客观有效的方式来比较产品,中介方式的成本高且效率低,急缺搜索和推荐信息平台。

13.5.5 应用提示

在“互联网+”作为国家战略、行业趋势和潮流的同时,把互联网金融以“三驾马车”之一的角色纳入国家战略,原因或许在于(不限于):

- 1) 作为一种应用模式,互联网金融在丰富和繁荣了金融市场、融合和再造了新型金融产品的同时,倒逼金融行业的回归和提升。
- 2) 作为一种营销渠道,互联网金融在拉近和强化了融贷关系、颠覆和重生了传统金融机构的同时,坐实小微企业的公平与机会。
- 3) 作为一种立国之道,互联网金融在纯粹和优化了金融环境、催生和打磨了普世金融服务的同时,助力个人主体的富民与普惠。

当然,上述目标和愿景的达成需要多边力量的共同努力,就计算机科学与技术侧而言,针对互联网金融涉及的各个环节提供技术服务、进行技术攻关从而助力细分目标的达成,应当是每一位计算机工作者的使命,也是机遇。作为一种应用示例,下文介绍互联网金融领域的一个细分应用场景:银行贷后风险评估。

作为一个传统金融机构，贷款是银行的重要获益来源。为了确保获益的稳定，风险控制是银行的一个重要核心业务，贷前、贷中、贷后分别有不同的业务系统（甚至部门）加以响应。比如贷前的业务流有贷前营销（银行端，期望找出优质客户）、贷款申请（用户端，贷款业务开始）、审查审批（银行端，资信评估）等；贷中一般是法律框架下的签约、放款等；贷后的业务流则有贷后检查、催收管理等。总体上来看，贷前的授信评估是银行风险控制的重中之重：让有资信的人（单位）获得贷款，这涉及征信的事务。事实上，经过多年的积淀，银行体系内已经储备了相对完备的征信模式，多年的实践也印证了这些征信手段的可信和可靠。不过潜在的一系列问题有：

1) 经过授信系统评估合格的企业获得款项后一定能够确保银行的投资收益吗？这是银行贷后一个“痛点”。

2) 是否存在放贷业务员出于绩效的压力，协助贷款企业（个人）利用征信系统的漏洞，使得并不具备相应资信的企业通过系统的资信评估？

3) 随着市场行情的变动，用于贷款的抵押物是否发生价值变动？这其实也是贷前信用评估的一个“痛点”。更进一步，此抵押物是否仅在当前一家银行抵押？

4) 随着市场行情的变动，贷款企业贷款预期的项目是否正常执行或者企业的运营状况是否与贷款授信时一致？

事实上，为了解决上述问题，贷后监管从来都是银行重中之重的风险监管环节，比如通过持续的信用评估机制对贷款企业进行持续的评级、比如通过实时的价格趋势分析对抵押物进行价值评估，并将上述所有活动与风险预警进行联动，贷后风险的评估一般是基于内部信息渠道的，比如嫁接（复用）贷前的授信评估体系。事实上，摆在银行（或者为银行贷后风险评估的开发商或服务商）面前用于授信评估的数据量已经足够巨大，且数据质量（源）往往有足够保障，但是存在的主要问题有：数据获取滞后现象严重，数据活跃度差；数据量虽然已经很大，但数据厚度差；数据大多集中在金融活动相关的数据，数据的广度差……

由于互联网的迅猛发展，出于政府公开、新闻报道、企业宣传等原因，互联网上积淀了大量反映企业运行状况的实时、动态信息，如果这部分信息得以利用，可以为银行的贷后风险做很好的补充（事实上，这部分信息对企业征信业大有裨益，那属于征信范畴，此处不议）。基于这样朴素的考虑，目前将互联网情报用于辅助贷后风险成了一个业界认可的趋势。按照大数据的流程和思路，核心的要点有：

1) 数据源配置：由于开展此类大数据分析具有显式的目的，因此数据源的配置往往是由现场工程师和业内专家进行有针对性的筛选，一般不可回避的数据源包括法院网站数据（尤其是开庭公告，这反映目标公司是否牵涉在某个官司中，以及是什么类型的官司，借此评估是否需要银行相关部门做止损操作）、新闻网站数据（尤其是财经频道数据，这会涉及目标企业涉及的新闻事件以及目标企业所在的行业的业界新闻等）、微博数据（一般微博大V发表的或者企业官方号发表的相关信息）、微信公众号（一般关注主流的财经、金融类公众号等）、地方论坛网站（反映老百姓的舆情以及老百姓“道听途说”的新闻）、人才招聘网站（反映

企业人才输入取向)等,此处无法罗列完备的数据源清单,而且在实际的运维过程中,还要允许现场工作人员动态地增删改相应的网站清单,以保证数据源质量的同时保证数据源的广度和数据的活度。

2) 数据标签化:采集到的数据经过清洗(比如过滤、去重)后,下面要进行的的就是标签化操作,因为所有网站数据能够反映某公司的风险预警输出应该是“A公司存在风险,具体风险是B”(这里的A指的是公司名,B指的是经由现场工作人员和行业专家配置的风险关键字,如撤资、起诉)。可以发现:此处的A和B都是某个网页数据的标签。因此,对采集的网页数据进行的操作就是利用配置好的热词为每个网页进行标签化。特别需要注意的是:这里的标签除了专家配置定义的关键词或者热词外,也包括反映网络舆情信息的情感信息;另外,由于各个公司可能存在异名的问题,这意味着在进行热词配置的时候,应该将可能的异名配置全;第三,在许多的新闻事件中,出现的主体往往是公司法人代表、股东等,这意味着需要在数据源收集的时候,把工商数据纳入数据源中,借此实现公司法人和自然人信息的关联。

3) 风险预警建模:仅仅对网页标签化还不足以反映企业的风险,必须对上述的标签建立关联。这是因为一般的风险预警都是基于一条风险规则的匹配进行的,以一个简单的风险规则“A公司,董事长撤资”为例,在标签化中得到的可能是:公司名A、董事长、撤资,唯有将这三个标签词在上下文预警上完全匹配,才是一个针对A公司的完整预警。另外,也有从机器学习的角度进行风险建模的,即将整个网页数据作为属性,通过学习,自动建立风险和网页的关系。

事实上,一个完备的互联网金融情报分析系统远不止上述的三个要点,至少还有如何高效数据采集、如何存取、如何提供服务(运维方式)等,此处无法一一罗列。但是,上面的简单分析至少表明了一个事实:在面向银行贷后风险辅助决策的应用目标下,开展系统设计与实现时,注重数据的混杂性,在注重数据的交叉复用、注重数据的相关性分析、注重数据的广度的同时,注重数据(源)的质量等,所有这些均是大数据思维。

13.6 本章小结

“政产学研商用”各界都对“互联网+”给予充分的关注和期待:从政府视角来看,“互联网+”是一种国家意志及国家战略,是推动产业转型升级、提升创新力和生产力的重要路径;从业界视角来看,“互联网+”是一种产业形态、一种应用模式或者一种商业模式;从研究者的视角来看,“互联网+”是一系列关键技术支撑下的应用,关键技术至少包括(移动)互联网、云计算、大数据。

而对普通老百姓而言,“互联网+”意味着一系列的普惠:

1) “互联网+工业”解放了纯粹的体力和一定脑力劳动付出的同时,提供灵活多样的职业路径,让人们的工作生涯更长、并且保持生产能力,而CPS支撑下的工作组织模式更加灵活,可以让员工在工作与私人生活之间以及个人发展与持续的职业发展之间实现更好的平衡。

2) “互联网+金融”能够让普通百姓(更包含小微企业)获得更公平和实惠的投融资机会。

3) “互联网+电子商务”让老百姓有更多的购买选择空间和更便捷的购买方式。

本章详细介绍了“互联网+”的相关背景以及我国政府发布的“互联网+”国家战略,同时着重提出了电子商务、工业互联网和互联网金融这“三驾马车”的基本概念,并给出了一些大数据应用的场景示例和可能。2015年7月1日,国务院公开发布的《关于积极推进“互联网+”行动的指导意见》提出“把互联网的创新成果与经济社会各领域深度融合,推动技术进步、效率提升和组织变革,提升实体经济的创新力和生产力,形成更广泛的以互联网为基础设施和实施工具的经济发展新形态”。在前述“三驾马车”的基础上又提出了进行“互联网+”行动的重点领域和行业,分别是:现代农业、益民服务、智慧能源、绿色生态、高效物流、便捷交通、人工智能、创业创新,详细的意见内容此处不赘述。

事实上,三百六十行都可以基于互联网思维进行互联网化,实现“互联网+”。比如:“互联网+医疗”就有了“春雨医生”“挂号网”等;“互联网+交通”就有了“滴滴打车”“e代驾”等;“互联网+电视”就有了“腾讯视频”“爱奇艺”等;“互联网+餐饮”就有了“饿了么”“美团外卖”等。虽然无法完全罗列所有的“互联网+”产品,但是可以看到的一个规律是,大部分的“互联网+”涉足的行业集中在第三产业,第二产业和第一产业的“互联网+”产品很少,其中的原因或许在于(不限于):

1) 老百姓接触的多是第三产业,自然对于第三产业的“互联网+”产品比较熟悉。

2) 第三产业涉及的人群基数大,商家进行“互联网+”的创新热情高。

3) 企业(行业)越轻越适合互联网化,创新“互联网+”产品,比如信息传输、软件和信息技术服务业、金融业、租赁和商务服务业、科学研究和技术服务业、教育、卫生和社会工作、文化、体育和娱乐业,这都属于第三产业。

4) 从产品的角度而言,如果产品本身不受时空限制,比如虚拟产品(诸如新闻、音乐、培训、金融服务、信息传输),则适合互联网改造。而如果产品本身是受时空限制的,“互联网+”的可能或许只在于通过对流程的优化和全链整合,极大缩短时空限制。

5) 对于更重的行业(第一产业、第二产业以及类似于物流这样的第三产业等),“互联网+”的贡献可能在于改造业务中的某些环节,比如通过互联网实现线上线下闭环;或者部分采用传统模式、部分进行互联网化改造,这点在制造业和物流业尤为明显。

本质上来看,无论是对相对轻的行业进行互联网化的“互联网+X”还是针对相对重的行业进行互联网优化的“X+互联网”,“互联网+”都是一种思维方式。将互联网作为一个技术手段、沟通渠道、产品形态、服务方式,以基因的方式植入目标产品或生产流程中,最终达到“of the people, by the people, for the people”的产品境界。如何达到这样的境界,需要多边智慧。不过在这个过程中,“互联网+”驱动和沉淀的大量数据(或许存放在互联网中,或许存放在企业数据库中,或者存在私人专有的设备中)势必会成为政产学研商用各界角力的重心。“大数据,大有所为”,这是一定的,但需要智慧。

毛泽东词作《忆秦娥·娄山关》云：“……雄关漫道真如铁，而今迈步从头越……”^① 总之，“互联网+”的风口来了。

本章参考文献

- [1] 马化腾. 互联网+国家战略行动路线图 [M]. 北京: 中信出版社, 2015.
- [2] 汤浔芳. 颠覆金融 [M]. 北京: 企业管理出版社, 2014.
- [3] 王吉斌, 彭盾. 互联网+: 传统企业的自我颠覆、组织重构、管理进化与互联网转型 [M]. 北京: 机械工业出版社, 2015.
- [4] 乌尔里希·森德勒. 工业 4.0 即将来袭的第四次工业革命 [M]. 邓敏, 李现民, 译. 北京: 机械工业出版社, 2015.
- [5] 新浪财经. 互联网金融 [M]. 北京: 东方出版社, 2014.

13.6 本章小结

一路求学，奔二地三校，缘因志在四方，五蕴皆空修六道轮回，但求七窍玲珑，八方风雨度数九寒冬，悄然大学十年。

大学十年，我分别就读于天津大学、河海大学和南京大学。因为留恋南京这座城市及感念于在此期间遇到的很多人，包括我的导师、前辈、同事和朋友，博士毕业后，在导师陈世福先生的鼓励和支持下，我很荣幸地成为南京大学的一名全职教师并任职至今。留校伊始，导师就建议我围绕智能信息处理展开相关科研工作，也正是因为这样的建议，我在这之后的许多工作都是围绕“数据”展开的，期间虽然也主持和承担了很多其他类型的项目，不过以数据驱动的研究一直在持续进行，所有的这些工作都成为本书成稿的最原始素材。

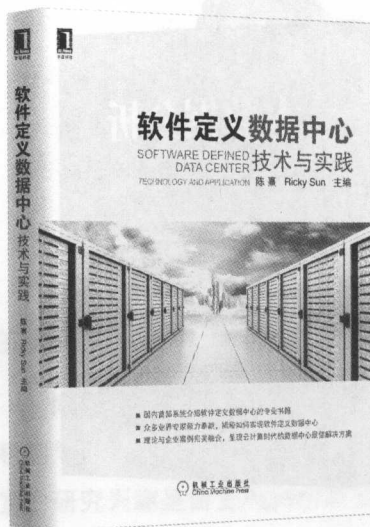
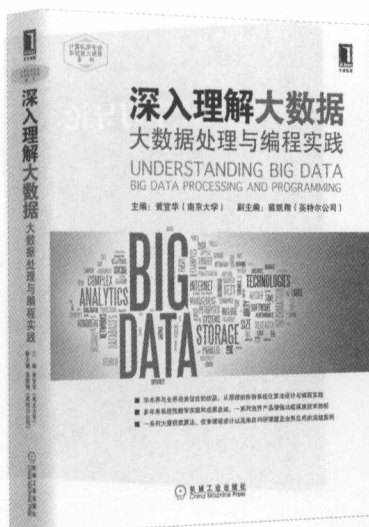
在本书的写作过程中，我除了参阅各类论文、书籍以及通过互联网获得各类电子文档和资料，也与很多同事、同行以及项目合作单位的朋友、本课题组的小伙伴们进行了交流。各种渠道富集的原始材料都为本书的成文和丰满奠定了最扎实的基础。当然，过往的申请书、投标书、演示文稿、教学课件、会议纪要等也是本书的重要参考资料，在此表示感谢。

毋庸置疑，全书的写作过程也是本人知识结构重新梳理的过程，因此我也在不断地回忆和复盘本人在课堂教学以及项目研发中的种种过往，并思考更加合理的迭代演进。这个过程是孤独的，孤独是因为我希望以更为独立、更为客观的第三方去慎思；但无疑这个过程也是温暖的，温暖是因为每当我感到百无聊赖几乎要放弃的时候，总有热心的前辈、同仁和小伙伴给予我莫大的关怀和支持。

本书行文即将完稿的时候，我和一个朋友说，我以这样的方式组织这样的内容，会得到读者的认同吗？这位朋友和我说，用心所致、言之有物、开卷有益、必有知音。我知道，这是鼓励，也是鞭策。因此，在本书再整理、再梳理、再润色的过程中，本人一直谨小慎微，努力行事，若本书能够给读者一些提示和借鉴，我将倍感荣幸。

在本书的行文过程中，我参阅了大量不同来源的参考文献，在此向这些参考文献的作者表示感谢，因为许多原因，文献标注可能有所遗漏，在此本人向这些作者表示歉意，请将及时知会我，我将在下一版本中予以修正。最后，再次感谢南京大学陈振宇教授、刘嘉副教

推荐阅读



深入理解大数据：大数据处理与编程实践

作者：黄宜华 等 ISBN：978-7-111-47325-1 定价：79.00元

本书在总结多年来MapReduce并行处理技术课程教学经验和成果的基础上，与业界著名企业Intel公司的大数据技术和产品开发团队和资深工程师联合，以学术界的教学成果与业界高水平系统研发经验完美结合，在理论联系实际的基础上，在基础理论原理、实际算法设计方法以及业界深度技术三个层面上，精心组织材料编写而成。

作为国内第一本经过多年课堂教学实践总结而成的大数据并行处理和编程技术书籍，本书全面地介绍了大数据处理相关的基本概念和原理，着重讲述了Hadoop MapReduce大数据处理系统的组成结构、工作原理和编程模型，分析了基于MapReduce的各种大数据并行处理算法和程序设计的思想方法。适合高等院校作为MapReduce大数据并行处理技术课程的教材，同时也很适合作为大数据处理应用开发和编程专业技术人员的参考手册。

—— 中国工程院院士、中国计算机学会大数据专家委员会主任 李国杰

软件定义数据中心——技术与实践

作者：陈熹 孙宇熙 ISBN：978-7-111-48317-5 定价：69.00元

国内首部系统介绍软件定义数据中心的专业书籍。

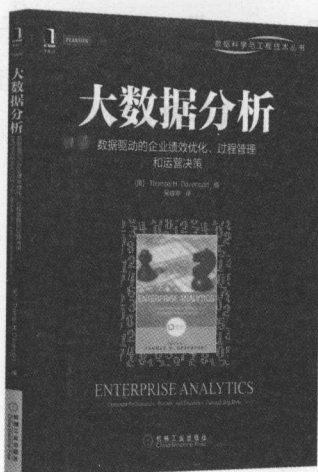
众多业界专家倾力奉献，揭秘如何实现软件定义数据中心。

理论与企业案例完美融合，呈现云计算时代的数据中心最佳解决方案。

有了以软件定义数据中心为基础的混合云，企业就可以进退有度，游刃有余，加上成功管理新的移动终端技术，可轻松进入“云移动”时代！这也是为什么软件定义数据中心最近获得大家注意的根本原因。EMC中国研究院编著的这本《软件定义数据中心：技术与实践》恰逢其时，它会给读者详细解说怎么实现软件定义数据中心。

—— VMware高级副总裁，EMC中国卓越研发集团创始人 Charles Fan

推荐阅读



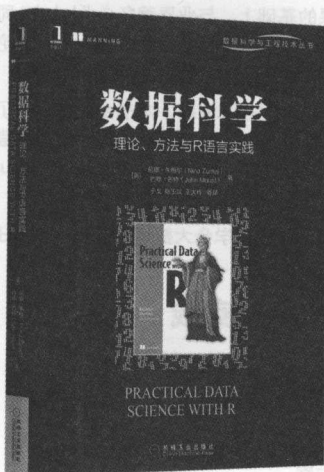
大数据分析：数据驱动的企业绩效优化、过程管理和运营决策

作者：Thomas H. Davenport ISBN: 978-7-111-49184-2 定价：59.00元



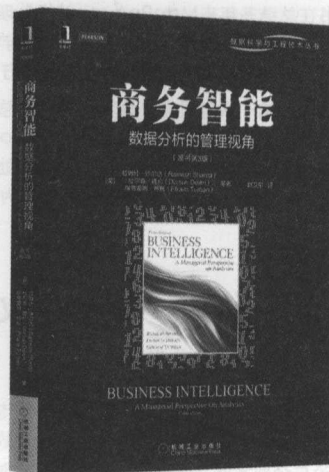
统计学习导论——基于R应用

作者：加雷斯·詹姆斯等 ISBN: 978-7-111-49771-4 定价：79.00元



数据科学：理论、方法与R语言实践

作者：尼娜·朱梅尔等 ISBN: 978-7-111-52926-2 定价：69.00元



商务智能：数据分析的管理视角（原书第3版）

作者：拉姆什·沙尔达等 ISBN: 978-7-111-49439-3 定价：69.00元

作者简介



王崇骏

博士、教授、博导，任职于南京大学计算机科学与技术系及软件新技术国家重点实验室，研究兴趣是自主Agent及多Agent系统、复杂网络理论及应用、大数据分析 & 智能系统。截至本书出版，主持和参与包括973、科技部重点专项、发改委专项、工信部产业化基金、国家自然科学基金、国家社会科学基金、省自然科学基金及支撑计划在内的国家及省部级基金与企事业资助项目50余项。在教育医疗类惠民行业、优政兴业类政府领域、互联网新经济领域有30余项科研成果获得产品化和商品化推广。



作者个人微信二维码

格物致知的技术手册
全周期描述大数据价值实现过程及相关理论与技术
慎思笃行的实施指南
多维度阐述大数据部署实施流程及应用攻略与提示
诚意正心的行动纲要
分层次论述大数据理性思维认知及多边挑战与机遇

这是一个最好的时代，也是一个最坏的时代；这是明智的时代，这是愚昧的时代；这是信任的纪元，这是怀疑的纪元；这是光明的季节，这是黑暗的季节；这是希望的春日，这是失望的冬日；我们面前应有尽有，我们面前一无所有；我们都将直上天堂，我们都将直下地狱……

狄更斯，《双城记》

我们很多人还没搞清楚什么是PC互联网的时候，移动互联来了；我们在没搞清楚移动互联的时候，大数据时代又来了……

马云，淘宝十周年演讲

大数据本非新鲜事，也非遥远事，更非神秘事……

引自本书

软件定义世界，数据驱动未来，未来已来……



投稿热线: (010) 88379604
客服热线: (010) 88379426 88361066
购书热线: (010) 68326294 88379649 68995259

华章网站: www.hzbook.com
网上购书: www.china-pub.com
数字阅读: www.hzmedia.com.cn

上架指导: 计算机/大数据

ISBN 978-7-111-54261-2



9 787111 542612 >

定价: 59.00元